An Oversampling Technique with Descriptive Statistics

HYONTAI SUG
Department of Computer Engineering,
Dongseo University,
47 Jurye-ro, Sasang-gu, Busan, 47011,
REPUBLIC OF KOREA

Abstract: - Oversampling is often applied as a means to win a better knowledge model. Several oversampling methods based on synthetic instances have been suggested, and SMOTE is one of the representative oversampling methods that can generate synthetic instances of a minor class. Until now, the oversampled data has been used conventionally to train machine learning models without statistical analysis, so it is not certain that the machine learning models will be fine for unseen cases in the future. However, because such synthetic data is different from the original data, we may wonder how much it resembles the original data so that the oversampled data is worth using to train machine learning models. For this purpose, I conducted this study on a representative dataset called wine data in the UCI machine learning repository, which is one of the datasets that has been experimented with by many researchers in research for knowledge discovery models. I generated synthetic data iteratively using SMOTE, and I compared the synthetic data with the original data of wine to see if it was statistically reliable using a box plot and t-test. Moreover, since training a machine learning model by supplying more high-quality training instances increases the probability of obtaining a machine learning model with higher accuracy, it was also checked whether a better machine learning model of random forests can be obtained by generating much more synthetic data than the original data and using it for training the random forests. The results of the experiment showed that small-scale oversampling produced synthetic data with statistical characteristics that were statistically slightly different from the original data, but when the oversampling rate was relatively high, it was possible to generate data with statistical characteristics similar to the original data, in other words, after generating high-quality training data, and by using it to train the random forests, it was possible to generate random forests with higher accuracy than using the original data alone, from 97.75% to 100%. Therefore, by supplying additional statistically reliable synthetic data as a way of oversampling, it was possible to create a machine-learning model with a higher predictive rate.

Key-Words: - Machine learning, classification, numerical attributes, oversampling, wine data, preprocessing, box plot, t-test.

Received: July 24, 2023. Revised: May 11, 2024. Accepted: June 24, 2024. Published: July 17, 2024.

1 Introduction

Discovering knowledge models with high accuracy on a given dataset is one of the most important issues in the field of machine learning and data mining. Since the accuracy of a machine learning algorithm can vary not only on the data given but also on how the algorithm processes the data, many machine learning algorithms have been proposed, and many experimental data have been open to the public in various areas, [1]. Because the best machine learning algorithm can vary from data to data, so, to encourage researchers to find the best machine learning algorithm for each data set, there are several public data sites, and the UCI machine learning repository, [2], is one of the most prominent sites. The site has several data sets and related research results that have been done using those data sets. Among the data in the UCI machine learning repository, the wine data attracted particular attention because we can achieve the high accuracy of several machine learning models even though the size of the data is small. In other words, the quality of the data is good. For example, various machine learning models, like logistic regression, neural networks, support vector machine, xgboost, and random forests, have been proposed for the wine data, and most of them have been reported to have a high accuracy of 95%±α, [3]. On the other hand, according to experiments, some machine learning algorithms do not have a high accuracy for the data. For example, locally weighted learning (LWL) has an accuracy of 88.8%, while zeroR has a low accuracy of 39.9% in 10-fold cross-validation. If the quality of the data is good and the

machine learning algorithm applied is suitable for the data, then we can expect the oversampled data to be good also by applying oversampling to the data, but we do not know how and how much oversampling should be done.

Meanwhile, it is true that the larger the data used for training, the higher the accuracy of machine learning algorithms. Of course, the data supplied for this must be of good quality and may represent the entire population well. To obtain more training data, fields such as image recognition make it easier for humans to label new image data that can be added to train machine learning models because it makes it easier for humans to recognize what an object is, [4], while data from science and social experiments rely on statistical methods because it is difficult for humans to evaluate the data.

In addition, in classification problems, when the number of data instances differs by class, it is common to apply oversampling to ensure that minority classes are not discriminated against classification. Now, considering that oversampling can generate more synthetic data than the original data, we want to see if feeding a lot of synthetic data into the training of a machine learning model can lead to a better machine learning model. Of course, we want to conduct a statistical analysis of how statistically reliable such training data is, because until now, the oversampled data has been used conventionally to train machine learning models without statistical analysis, so it is not certain that the machine learning models will be fine for unseen cases in the future.

From now on section 2 covers related work, section 3 deals with problem formation, section 4 covers experimentation, and section 5 presents the conclusion.

2 Related Work

Wine is one of the most consumed alcoholic beverages, and as a result, the analysis of wine-related public data has attracted the attention of many researchers. For example, there are two kinds of wine data on the UCI machine learning repository, one called ‘wine’ and the other called ‘wine quality’. Of these two types of data, we note that it is worth oversampling for ‘wine’ data that is relatively small-sized, yet known to have highly accurate knowledge models. The data set contains 178 instances consisting of 13 chemical constituents to classify three wine cultivars from Italy. Since wine data was loaded into the UCI machine learning repository in 1991, a lot of research has been conducted.

Statistical inference has been studied for variables of interest when auxiliary variables are observed along with the variables of interest that do not have enough data, and an experiment was done using the wine data, [5]. More recently, detailed analysis for each 13 attributes of the wine data was given, and neural networks and support vector machine classifiers were made to generate accurate classifiers achieving the accuracy of 94.4% ~ 97.8% in 5-fold cross-validation, [6]. An ensemble-based method called the ensemble learning method for spectral clustering reported achieving a performance of 75.8% in clustering, [7]. Graph neural networks achieved an accuracy of 97.5% to 98% in 10-fold cross-validation for the data set, [8]. The multi-output neural tree that has a tree structure whose nodes are neural achieved an error rate of 15% to 16% where 20% of the data is used for testing, [9].

When we do not have sufficient sized data, we may rely on oversampling. Simple oversampling oversamples a specific class to give more attention to training a machine learning model, [10]. Because simple oversampling increases the likelihood of overfitting by introducing replicated samples, oversampling based on synthetic samples was invented, [11]. SMOTE is one of the representative oversampling methods that generate synthetic instances of a minor class. SMOTE stands for Synthetic Minority Over-sampling TEchnique. It selects k-nearest neighbors and generates a synthetic data instance by multiplying an interval value of continuous attributes of the k-nearest neighbors with a random number between 0 and 1, [12]. There are several variant algorithms of SMOTE. Among them, borderline-SMOTE and ADSYN (Adaptive Synthetic Sampling Approach) may be two representatives. Borderline-SMOTE uses a synthetic data generation oversampling method that focuses on the hard-to-classify parts of the sample by increasing the number of samples that are borderline with other major classes, [13]. ADASYN aims to improve the accuracy of classification by generating synthetic data for minority classes, which are easy to misclassify because their values are similar to those of major classes, [14]. On the other hand, an algorithm called SNOCC that creates multiple clusters for each major and minority class and oversamples each cluster belonging to the minority class so that it will become similar to the number of instances belonging to the majority class has been suggested, [15] and argued that the synthetic data generated by SNOCC shows better statistical characteristics than the synthetic data generated by SMOTE by considering the entire neighborhood in

the cluster while SMOTE considers a fixed number of the nearest neighbors, k.

Random forests generate many random decision trees based on random sampling with replacement and use the many decision trees to classify, [16]. When each decision tree is generated, a random selection of root attributes for each subtree and no pruning is performed, and classes are determined by majority vote by the decision trees in the forest so that overfitting can be avoided. The most important property of random forests is that it has randomness when creating the trees that make up the forest so that the prediction of each tree is de-correlated with each other, and as a result, the generalization performance is improved. Random forests are known to be one of the most reasonable machine learning algorithms across a wide range of data, [17]. Other factors that affect the performance of random forests are the size and dimensionality of the sample. Because a smaller sample size may not represent the population well, the sample size may affect the performance of random forests, [18] and the dimensionality of the data affects the size of the sample required to generate good random forests, [19].

3 Problem Formulation

The wine data set has thirteen continuous or numerical conditional attributes and one decisional attribute that classifies wine quality into three classes. The data set is the result of a chemical analysis of three different wines of class 1, 2, and 3 grown in the same region in Italy. According to the data description in the UCI machine learning repository, [3], several machine learning algorithms like logistic regression, neural network, support vector, xgboost, and random forests showed high accuracy in classification.

On the other hand, the method of applying oversampling as a method to improve the knowledge model has been studied, and there are two methods of oversampling; Simple oversampling and synthetic data-based oversampling like SMOTE. In this paper, we want to apply SMOTE to generate new synthetic data to avoid overfitting, and by identifying the statistical properties of the new synthetic data concerning the original data, we aim to confirm the quality of the data.

Since the accuracy of the machine learning model can be improved by supplying high-quality training data, we would like to see whether the supply of a large amount of high-quality synthetic data through oversampling can contribute to the improvement of the accuracy of the machine learning models. As a means of measuring the quality of the data, we want to use a box plot that makes it easy to see the mean, median, and quartiles as well as outliers by eye. In addition, we want to use a t-test to confirm whether the newly created data and the original data statistically belong to the same population.

The t-test is a statistical method used to compare whether the difference in the mean values between two groups is statistically the same or different. In other words, in a two-sample t-test, the t-value is the mean difference between the two groups divided by the mean standard error. Before the t-test, the equal variance test will be performed first, and the reason for the equal variance test is that it can indirectly confirm whether the target of statistical analysis is extracted from the same population. Equal variance testing of Levene, which is also called Levene's F-test, is mainly used because it can be used even when there is no certainty that the data is in normal distribution, [20]. In the assumption of equal variance, if the significance level p-value is greater than 0.05, then equal variance can be assumed so that Student's t-test will be applied. If the p-value is less than or equal to 0.05, then equal variance cannot be assumed, so Welch's t-test will be applied. If we use IBM SPSS for the t-test, [21], the top line shows the t-test result when it is equal variance, and the bottom line shows the t-test result when it is not equal variance. Finally, in the t-test, the smaller the p-value, the more significant the difference between the two groups is, so the p-value of 0.05 or less is the criterion for such a judgment, [22].

For the box plots, MS Excel 2016 will be used, and for the t-test, a well-known tool, IBM SPSS, will be used for the experiment. For oversampling and to generate random forests an open-source tool called Weka will be used, [23].

3.1 Experimental Procedure

We want to check whether we can find random forests of high accuracy by adding synthetic data of oversampling for the wine dataset, and we repeat oversampling until the random forests are no longer improved. Oversampling of 100% on the classes with the highest number of misclassifications in the resulting confusion matrix will be repeated until there is no further improvement in misclassification. The misclassification will be judged by two factors. The first factor is the number of misclassifications in 10-fold CV(cross-validation) when we use the original data and oversampled data together for training and testing, and the second factor is the number of misclassifications of the random forests

that are trained by oversampled instances only and tested by the original data.

- When selecting a class for the next iteration, we prioritize the results of 10-fold CV, followed by the test results of the original data.
- If we do not have misclassifications in the 10-fold CV but still have misclassifications in the test result by the original data, we also consider them for further iteration.
- If the number of misclassifications in each class is identical in 10-fold CV or the test result by the original data, a class for the next iteration is determined based on the fact that one of them has different test results.
- If the number of misclassifications in each class is identical in 10-fold CV and the test result by the original data, we pick one of them randomly for the next iteration.

4 Experimentation

'Wine' data set is used for experiments in the UCI machine learning repository, [3]. The goal of experiments is to find the best models of random forests as accurately as possible based on progressive and repetitive oversampling and to perform statistical tests on the data.

4.1 Wine Dataset

The data set is the result of a chemical analysis of three different wines of class 1, 2, and 3 grown in the same region in Italy. The data set has 178 records and has 13 conditional attributes and one decisional attribute, named class. The 13 conditional attributes consist of numerical attributes as in Table 1.

Table 1. The Property of attributes of the wine data set

Attribute	Value Range	Distinct values	Mean	Standard Deviation
Alcohol	11.03 ~ 14.83	126	13.001	0.812
Malic acid	0.74 ~ 5.8	133	2.336	1.117
Ash	1.36 ~ 3.23	79	2.367	0.274
Alkalinity of Ash	10.6 ~ 30	63	19.495	3.34
Magnesium	70 ~ 162	53	99.742	14.282
Total phenols	0.98 ~ 3.88	97	2.295	0.626
Flavonoids	0.34 ~ 5.08	132	2.029	0.999
Nonflavonoid phenols	0.13 ~ 0.66	39	0.362	0.124
Proanthocyanins	0.41 ~ 3.58	101	1.591	0.1572
Color Intensity	1.28 ~ 13	132	5.058	2.318
Hue	0.48 ~ 1.71	78	0.957	0.229
OD280 or OD315 of diluted wines	1.27 ~ 4	122	2.612	0.71
Proline	278 ~ 1680	121	746.893	314.907
Class	3 class values (1, 2, 3)			

Malic acid has a strong link to wine taste. Ash is the inorganic matter that remains after evaporation and incineration of wine. The alkalinity of ash is a measure of weak alkalinity dissolved in water. Magnesium has its most significant function as an essential component of chlorophyll to the health of grapevines. Total phenols in wine are important in estimating the taste and health benefits of wine. Flavonoids are a type of antioxidant that is high in red wines. Nonflavonoid phenols can include several subclasses of importance to wine. Proanthocyanins are condensed tannins and the most abundant class of phenols in wine. OD280/OD315 of diluted wines determine the protein content of various wines. Proline is an amino acid in wine.

Table 2 shows random forests for the original data set. One thousand random trees are trained and tested with 10-fold cross-validation.

Table 2. The result of random forests for the original data of the wine data set

Accuracy in 10-CV (%)	Confusion matrix			No. of Misclassified
97.7528	58	1	0	4
	1	68	2	
	0	0	48	

Note that class 1 consists of 59 instances, class 2 consists of 71 instances, and class 3 consists of 48 instances.

4.1.1 Progressive and Repetitive Oversampling

Oversampling of 100% for the classes with the highest number of misclassifications in the confusion matrix was repeated until there was no further improvement in misclassification using SMOTE. Default parameters of SMOTE of k = 5 which is the number of nearest neighbors and seed = 1 which is the seed for a random number between 0 to 1 were used. The misclassification in 10-fold cross-validation is considered when original data and oversampled data are used together for training and testing, and also the number of misclassifications of the random forests that are trained by oversampled instances only and tested by the original data. We pick the class with the highest number of misclassifications for the next oversampling in the 10-fold CV. If the values are identical, we pick them randomly. If the misclassification in the test with the original data still needs to be improved, we do oversampling further. Table 3 and Table 4 show the results of the experiment, and Table 3 shows the result of the experiment using 10-fold cross-validation.

Table 3. The result of progressive oversampling with 10-CV for wine data

Iter-ation No.	Oversampled classes	Accuracy in 10-CV (%)	Confusion matrix			No. of Misclassified
1	2	98.3963	58	1	0	4
			1	139	2	
			0	0	48	
2	2	98.7212	55	4	0	5
			1	283	0	
			0	0	48	
3	1	99.5566	118	0	0	2
			1	283	0	
			0	1	47	
4	3	99.1968	117	1	0	4
			1	282	1	
			0	1	95	
5	2	99.7442	117	1	0	2
			1	567	0	
			0	0	96	
6	1	100	236	0	0	0
			0	568	0	
			0	0	96	
7	3	99.6988	236	0	0	3
			0	567	1	
			0	2	190	
8	3	99.8316	235	1	0	2
			1	567	0	
			0	0	384	
9	1	100	472	0	0	0
			0	568	0	
			0	0	384	

Table 4. The corresponding result of progressive oversampling when oversampled data are used for training and the original data are used for testing for the wine data set

Iter-ation No,	Over-sampled classes	Accuracy with the original data (%)	Confusion matrix			No. of Misclassified
1	2	39.8876	0	59		107
			0	71	0	
			0	48	0	
2	2	39.8876	0	59		107
			0	71	0	
			0	48	0	
3	1	70.7068	55	4	0	50
			0	71	0	
			2	46	47	
4	3	94.9438	54	5	0	9
			0	71	0	
			0	4	44	
5	2	94.9438	55	4	0	9
			0	71	0	
			0	5	43	
6	1	97.7528	58	1	0	4
			0	71	0	
			0	3	45	
7	3	98.3141	58	1	0	3
			0	71	0	
			0	2	46	
8	3	99.4382	58	1	0	1
			0	71	0	
			0	0	48	
9	1	100	59	0	0	0
			0	71	0	
			0	0	48	

Table 4 shows the corresponding result of the experiment when we use over-sampled instances only for training and the original data for testing.

Note that we do further oversampling after iteration 6 because the misclassification in the test with original data is still not satisfactory.

4.1.2 Statistical test for Oversampled Data of Wine

Statistical tests are done to see the properties of oversampled data for each attribute. Data at iterations 4 and 9 are tested because all three classes are oversampled for the first time at iteration 4 and iteration 9 is the final iteration.

4.1.2.1 Statistical Test for Attribute Alcohol

Figure 1 shows the 5 box plots for the attribute 'alcohol' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively. Note that the data of 10-fold CV contains the original data as well as oversampled data. We can see the final oversampling generated the data of a similar but slightly narrow box. The oversampling at iteration 4 generated lower Q1(the first quartile), lower Q2(the second quartile), lower Q3(the third quartile), and lower mean.

Fig. 1: Box plots for attribute 'alcohol' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively.

A t-test for the mean was carried out on the synthetic data for the attribute alcohol at iterations 4 and 9. From the test equal variance was assumed, because F=5.136 with significance=0.24 for the data at iteration 4 and F=3.307 with significance=0.69 for the data at iteration 9. Table 5 and Table 6 summarize the results. Table 5 shows the result of the t-test for the oversampled data at iteration 4. SD means standard deviation in the table.

Table 5. Result of the t-test for the oversampled data at iteration 4 for the attribute alcohol

Data at	Mean	SD	t	p
Original	13.0006	0.8118	4.084	0.0
Iteration 4	12.7120	0.7230		

Because p=0.0 which is less than 0.05, we can see that the difference in the means is statistically significant, so we can say that the oversampled data only for attribute alcohol are not good at iteration 4. Table 6 shows the result of the t-test for the oversampled data at iteration 9.

Table 6. Result of the t-test for the oversampled data at iteration 9 for the attribute alcohol

Data at	Mean	SD	t	p
Original	13.0006	0.8118	-0.297	0.766
Iteration 9	13.0183	0.7324		

Because p=0.766, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data for attribute alcohol are good at iteration 9.

4.1.2.2 Statistical Test for Attribute Malic Acid

Figure 2 shows the 5 box plots for the attribute 'malic acid'. We can see the final oversampling generated the data of a similar but slightly narrow box. But, the oversampling at iteration 4 generated lower Q3, lower mean, and lower upper bound.

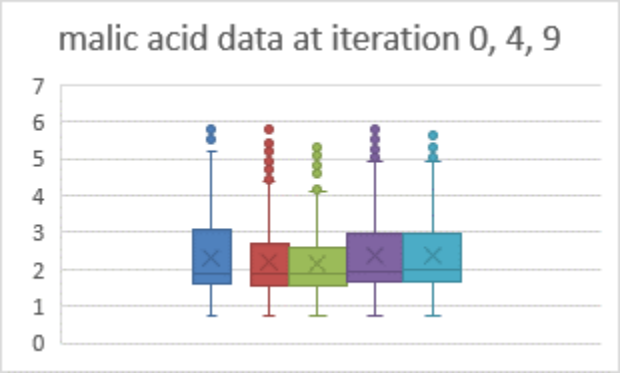

Fig. 2: Box plots for attribute 'malic acid' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively.

A t-test for the mean was carried out on the synthetic data for the attribute malic acid at iterations 4 and 9. Because F=13.421 with significance=0.0 for the data at iteration 4, we cannot assume equal variance for the original and oversampled data at iteration 4. Table 7 shows the result of the t-test for the oversampled data at iteration 4.

Table 7. Result of the t-test for the oversampled data at iteration 4 for the attribute malic acid

Data at	Mean	SD	t	p
Original	2.3363	1.1171	1.861	0.064
Iteration 4	2.1535	0.9198		

Because p=0.064, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute malic acid are good at iteration 4.

Because F=3.879 with significance=0.049 for the data at iteration 9, we cannot assume equal variance. Table 8 shows the result of the t-test for the oversampled data at iteration 9.

Table 8. Result of the t-test for the oversampled data at iteration 9 for the attribute malic acid

Data at	Mean	SD	t	p
Original	2.3363	1.1171	-1.051	0.294
Iteration 9	2.4300	0.9970		

Because p=0.294, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data for attribute malic acid are good at iteration 9.

4.1.2.3 Statistical Test for Attribute Ash

Figure 3 shows the 5 box plots for the attribute 'ash'. We can see the final oversampling generated the data of a similar but slightly narrow box as we can see the Q1 and Q3. The oversampling at iteration 4 generated lower Q1, lower Q2, lower Q3, and lower mean.

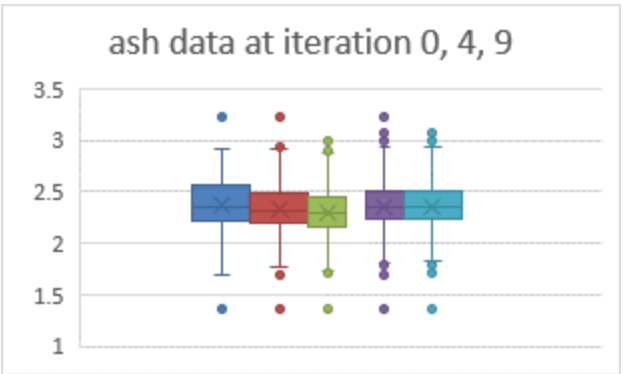

Fig. 3: Box plots for attribute 'ash' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively

A t-test for the mean was carried out on the synthetic data for the attribute ash at iterations 4 and 9. From the test equal variance was assumed, because F=1.961 with significance=0.162 for the data at iteration 4. Table 9 shows the result of the t-test for the oversampled data at iteration 4.

Table 9. Result of the t-test for the oversampled data at iteration 4 for the attribute ash

Data at	Mean	SD	t	p
Original	2.3665	0.2743	2.675	0.008
Iteration 4	2.3022	0.2467		

Because p=0.008, which is less than 0.05, we can see that the difference in the means is statistically significant, so we can say that the oversampled data only for attribute ash are not good at iteration 4. Because F=8.253 with significance=0.009 for the data at iteration 9, we cannot assume equal variance. Table 10 shows the result of the t-test for the oversampled data at iteration 9.

Table 10. Result of the t-test for the oversampled data at iteration 9 for the attribute ash

Data at	Mean	SD	t	p
Original	2.3665	0.2743	0.301	0.763
Iteration 9	2.3600	0.2274		

Because p=0.763, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute ash are good at iteration 9.

4.1.2.4 Statistical Test for Attribute Alkalinity of Ash

Figure 4 shows the 5 box plots for the attribute 'alkalinity of ash'. We can see the final oversampling generated the data of a similar but slightly narrow box as we can see the Q1 and Q3. The oversampling at iteration 4 generated a larger Q1, and lower upper boundary, and a smaller lower boundary.

A t-test for the mean was carried out on the synthetic data for the attribute alkalinity of ash at iterations 4 and 9. From the test equal variance cannot be assumed, because F=11.666 with significance=0.001 for the data at iteration 4 and F=12.557 with significance=0.0 for the data at iteration 9. Table 11 and Table 12 summarize the results. Table 11 shows the result of the t-test for the oversampled data at iteration 4.

Fig. 4: Box plots for attribute 'alkalinity of ash' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively.

Table 11. Result of the t-test for the oversampled data at iteration 4 for the attribute alkalinity of ash

Data at	Mean	SD	t	p
Original	19.4949	3.3396	-1.095	0.275
Iteration 4	19.8115	2.5911		

Because p=0.275, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute alkalinity of ash are good at iteration 4. Table 12 shows the result of the t-test for the oversampled data at iteration 9.

Table 12. Result of the t-test for the oversampled data at iteration 9 for the attribute alkalinity of ash

Data at	Mean	SD	t	p
Original	19.4949	3.3396	-0.263	0.793
Iteration 9	19.5637	2.6764		

Because p=0.793, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute alkalinity of ash are good at iteration 9.

4.1.2.5 Statistical Test for Attribute Magnesium

Figure 5 shows the 5 box plots for the attribute 'magnesium'. We can see the final oversampling generated the data of a similar but slightly narrow box as we can see the Q1, Q3, upper, and lower boundary. The oversampling at iteration 4 generated higher Q1, lower Q2, lower Q3, lower mean, and lower upper boundary.

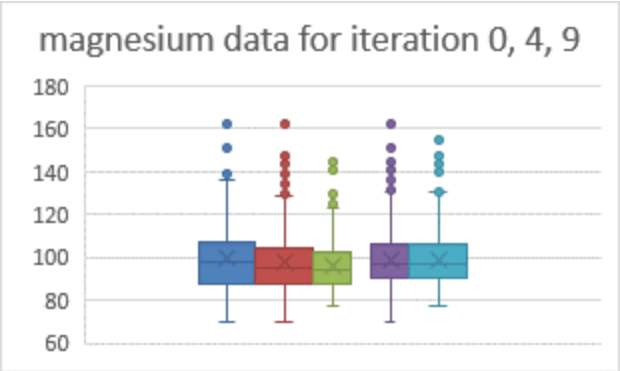

Fig. 5: Box plots for attribute 'magnesium' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively.

A t-test for the mean was carried out on the synthetic data for the attribute magnesium from iterations 4 and 9. From the test equal variance cannot be assumed, because F=6.185 with significance=0.013 for the data at iteration 4 and F=8.827 with significance=0.003 for the data at iteration 9. Table 13 and Table 14 summarize the results. Table 13 shows the result of the t-test for the oversampled data at iteration 4.

Table 13. Result of the t-test for the oversampled data at iteration 4 for the attribute magnesium

Data at	Mean	SD	t	p
Original	99.7416	14.2825	2.764	0.006
Iteration 4	96.2611	11.7675		

Because p=0.006, which is less than 0.05, we can see that the difference in the means is statistically significant, so we can say that the oversampled data only for attribute magnesium are not good at iteration 4. Table 14 shows the result of the t-test for the oversampled data at iteration 9.

Table 14. Result of the t-test for the oversampled data at iteration 9 for the attribute magnesium

Data at	Mean	SD	t	p
Original	99.7416	14.2825	0.749	0.454
Iteration 9	98.9025	11.5992		

Because p=0.454, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute magnesium are good at iteration 9.

4.1.2.6 Statistical Test for Attribute Total Phenols

Figure 6 shows the 5 box plots for attribute 'total phenols'. We can see the final oversampling generated the data of a similar but slightly narrow box as we can see the Q1, Q3, upper, and lower boundary. However, the oversampling at iteration 4 generated a larger Q1, lower Q2, and lower Q3.

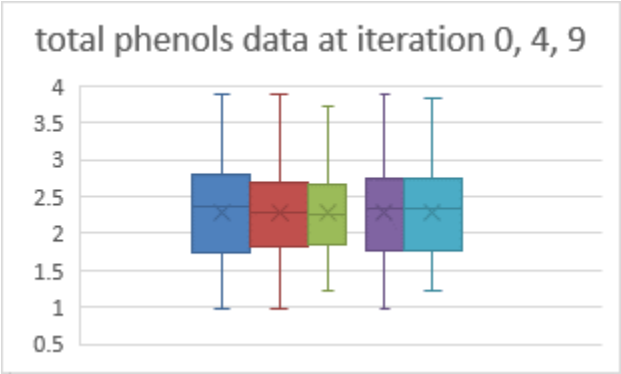

Fig. 6: Box plots for attribute 'total phenols' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively.

A t-test for the mean was carried out on the synthetic data for the attribute total phenols from iterations 4 and 9. From the test equal variance cannot be assumed, because F=10.431 with significance=0.001 for the data at iteration 4. Table 15 shows the result of the t-test for the oversampled data at iteration 4.

Table 15. Result of the t-test for the oversampled data at iteration 4 for the attribute total phenols

Data at	Mean	SD	t	p
Original	2.2951	0.6259	0.283	0.777
Iteration 4	2.2794	0.5339		

Because p=0.777, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute total phenols are good at iteration 4. Because F=1.987 with significance=0.159 for the data at iteration 9, we can assume equal variance. Table 16 shows the result of the t-test for the oversampled data at iteration 9.

Table 16. Result of the t-test for the oversampled data at iteration 9 for the attribute total phenols

Data at	Mean	SD	t	p
Original	2.2951	0.6259	-0.069	0.945
Iteration 9	2.2918	0.5825		

Because p=0.945, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute total phenols are good at iteration 9.

4.1.2.7 Statistical Test for Attribute Flavonoids

Figure 7 shows the 5 box plots for the attribute 'flavonoids'. We can see the final oversampling generated the data of a similar but slightly narrow box as we can see the Q1, Q3, upper, and lower boundary. But, the oversampling at iteration 4 generated larger Q1, lower Q2, lower Q3, lower mean and lower upper boundary.

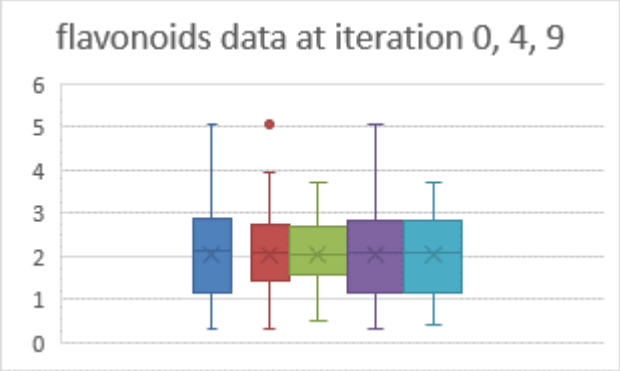

Fig. 7: Box plots for attribute 'flavonoids' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively.

A t-test for the mean was carried out on the synthetic data for the attribute flavonoids from iterations 4 and 9. From the test equal variance cannot be assumed, because F=26.164 with significance=0.0 for the data at iteration 4. Table 17 shows the result of the t-test for the oversampled data at iteration 4.

Table 17. Result of the t-test for the oversampled data at iteration 4 for the attribute flavonoids

Data at	Mean	SD	t	p
Original	2.0293	0.9989	-0.147	0.883
Iteration 4	2.0420	0.7753		

Because p=0.883, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute flavonoids are good at iteration 4. Because F=1.06 with significance=0.303 for the data at iteration 9, we can assume equal variance. Table 18 shows the result of the t-test for the oversampled data at iteration 9.

Table 18. Result of the t-test for the oversampled data at iteration 9 for the attribute flavonoids

Data at	Mean	SD	t	p
Original	2.0293	0.9989	0.102	0.919
Iteration 9	2.0215	0.9394		

Because p=0.919, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute flavonoids are good at iteration 9.

4.1.2.8 Statistical Test for Attribute Nonflavonoid Phenols

Figure 8 shows the 5 box plots for the attribute 'nonflavonoid phenols'. We can see the final oversampling generated the data of a similar but slightly narrow box as we can see the Q1 and Q3. However, the oversampling at iteration 4 generated larger Q1, larger Q2, and lower Q3.

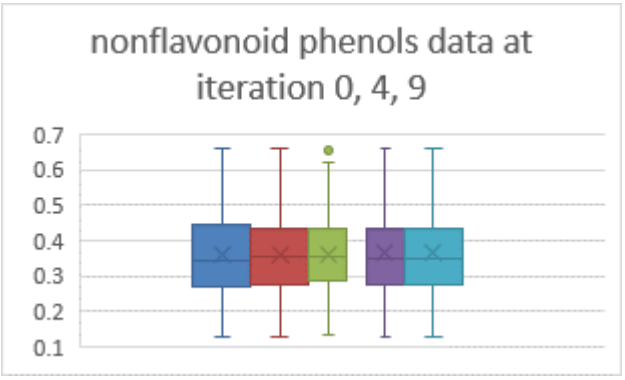

Fig. 8: Box plots for attribute 'nonflavonoid phenols' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively.

A t-test for the mean was carried out on the synthetic data for the attribute nonflavonoid phenols at iterations 4 and 9. From the test equal variance cannot be assumed, because F=14.807 with significance=0.0 for the data at iteration 4. Table 19 shows the result of the t-test for the oversampled data at iteration 4.

Table 19. Result of the t-test for the oversampled data at iteration 4 for the attribute nonflavonoid phenols

Data at	Mean	SD	t	p
Original	0.3619	0.1245	-0.046	0.963
Iteration 4	0.3624	0.1022		

Because p=0.963, which is greater than 0.05, we can see that the difference in the means is

statistically insignificant, we can say that the oversampled data only for attribute nonflavonoid phenols are good at iteration 4. Because F=12.677 with significance=0.0 for the data at iteration 9, we cannot assume equal variance. Table 20 shows the result of the t-test for the oversampled data at iteration 9.

Table 20. Result of the t-test for the oversampled data at iteration 9 for the attribute nonflavonoid phenols

Data at	Mean	SD	t	p
Original	0.3619	0.1245	-0.113	0.910
Iteration 9	0.3630	0.1054		

Because p=0.910, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute nonflavonoid phenols are good at iteration 9.

4.1.2.9 Statistical Test for Attribute Proanthocyanins

Figure 9 shows the 5 box plots for the attribute 'proanthocyanins'. We can see the final oversampling generated the data of a similar box. The oversampling at iteration 4 generated larger Q1, lower Q3, and smaller upper boundary.

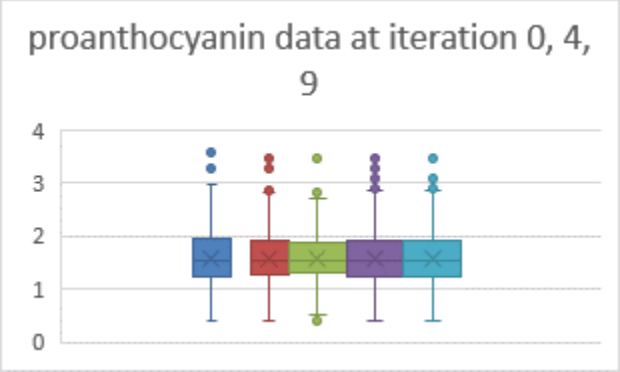

Fig. 9: Box plots for attribute 'proanthocyanins' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively.

A t-test for the mean was carried out on the synthetic data for the attribute proanthocyanins from iterations 4 and 9. From the test equal variance cannot be assumed, because F=5.056 with significance=0.025 for the data at iteration 4. Table 21 shows the result of the t-test for the oversampled data at iteration 4.

Table 21. Result of the t-test for the oversampled data at iteration 4 for the attribute proanthocyanins

Data at	Mean	SD	t	p
Original	1.5909	0.5724	0.123	0.902
Iteration 4	1.5846	0.4906		

Because p=0.902, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute proanthocyanins are good at iteration 4. Because F=5.505 with significance=0.019 for the data at iteration 9, we cannot assume equal variance. Table 22 shows the result of the t-test for the oversampled data at iteration 9.

Table 22. Result of the t-test for the oversampled data at iteration 9 for the attribute proanthocyanins

Data at	Mean	SD	t	p
Original	1.5909	0.5724	-0.462	0.645
Iteration 9	1.5701	0.4926		

Because p=0.645, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute proanthocyanins are good at iteration 9.

4.1.2.10 Statistical Test for Attribute Color Intensity

Figure 10 shows the 5 box plots for the attribute 'color intensity'. We can see the final oversampling generated the data of a similar box. The oversampling at iteration 4 generated lower Q1, lower Q2, lower Q3, lower mean, smaller lower boundary, and smaller upper boundary.

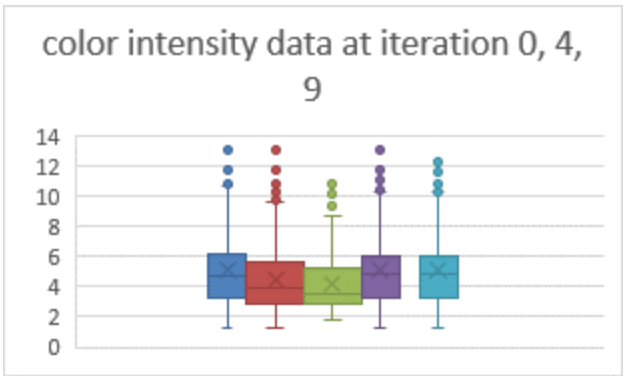

Fig. 10: Box plots for attribute 'color intensity' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively

A t-test for the mean was carried out on the synthetic data for the attribute color intensity from iterations 4 and 9. From the test equal variance cannot be assumed, because F=9.129 with significance=0.003 for iteration 4. Table 23 shows the result of the t-test for the oversampled data at iteration 4.

Table 23. Result of the t-test for the oversampled data at iteration 4 for the attribute color intensity

Data at	Mean	SD	t	p
Original	5.0581	2.3183	4.202	0.0
Iteration 4	4.2042	1.8848		

Because p=0.0, which is less than 0.05, we can see that the difference in the means is statistically significant, so we can say that the oversampled data only for attribute color intensity are not good at iteration 4. Because F=2.529 with significance=0.112 for the data at iteration 9, we can assume equal variance. Table 24 shows the result of the t-test for the oversampled data at iteration 9.

Table 24. Result of the t-test for the oversampled data at iteration 9 for the attribute color intensity

Data at	Mean	SD	t	p
Original	5.0581	2.3183	0.045	0.964
Iteration 9	5.0504	2.0988		

Because p=0.964, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute color intensity are good at iteration 9.

4.1.2.11 Statistical Test for Attribute Hue

Figure 11 shows the 5 box plots for the attribute 'hue'. We can see the final oversampling generated the data of a similar but slightly narrow box as we can see the Q1 and Q3. The oversampling at iteration 4 generated a larger Q1, larger Q2, lower Q3, and larger mean.

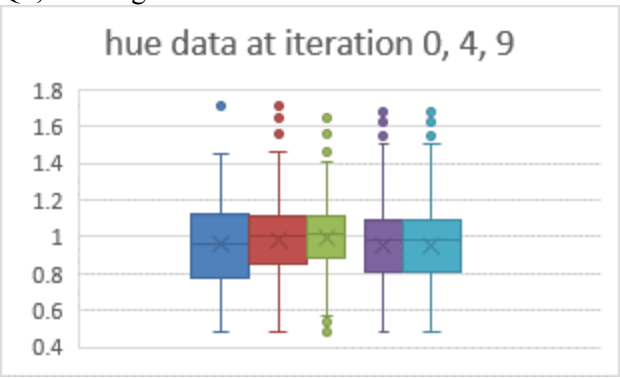

Fig. 11: Box plots for attribute 'hue' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively

A t-test for the mean was carried out on the synthetic data for the attribute hue from iterations 4 and 9. From the test equal variance cannot be assumed, because F=9.169 with significance=0.003 for the data at iteration 4. Table 25 shows the result of the t-test for the oversampled data at iteration 4.

Table 25. Result of the t-test for the oversampled data at iteration 4 for the attribute hue

Data at	Mean	SD	t	p
Original	0.9574	0.2286	-1.830	0.068
Iteration 4	0.9946	0.1956		

Because p=0.068, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute hue are good at iteration 4. Because F=4.804 with significance=0.029 for the data at iteration 9, we cannot assume equal variance. Table 26 shows the result of the t-test for the oversampled data at iteration 9.

Table 26. Result of the t-test for the oversampled data at iteration 9 for the attribute hue

Data at	Mean	SD	t	p
Original	0.9574	0.2286	0.334	0.739
Iteration 9	0.9514	0.2048		

Because p=0.739, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute hue are good at iteration 9.

4.1.2.12 Statistical Test for Attribute OD280 or OD315 of Diluted Wines

Figure 12 shows the 5 box plots for the attribute 'OD280 or OD315 of diluted wines'. We can see the final oversampling generated the data of a similar but slightly narrow box as we can see the Q1. But, the oversampling at iteration 4 generated larger Q1, and smaller Q3.

A t-test for the mean was carried out on the synthetic data for the attribute OD280 or OD315 of diluted wines from iterations 4 and 9. From the test equal variance cannot be assumed, because F=28.069 with significance=0.0 for the data at iteration 4. Table 27 shows the result of the t-test for the oversampled data at iteration 4.

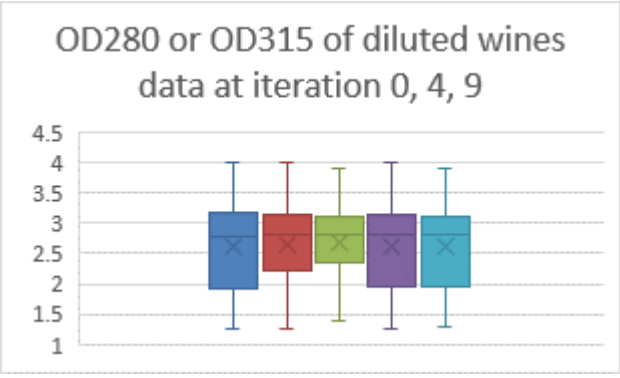

Fig. 12: Box plots for attribute 'OD280 or OD315 of diluted wines' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively

Table 27. Result of the t-test for the oversampled data at iteration 4 for the attribute OD280 or OD315 of diluted wines

Data at	Mean	SD	t	p
Original	2.6117	0.7100	-1.127	0.261
Iteration 4	2.6809	0.5481		

Because p=0.261, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute OD280 or OD315 of diluted wines are good at iteration 4. Because F=4.002 with significance=0.046 for the data at iteration 9, we cannot assume equal variance. Table 28 shows the result of the t-test for the oversampled data at iteration 9.

Table 28. Result of the t-test for the oversampled data at iteration 9 for the attribute OD280 or OD315 of diluted wines

Data at	Mean	SD	t	p
Original	2.6117	0.7100	0.148	0.883
Iteration 9	2.6034	0.6461		

Because p=0.883, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute OD280 or OD315 of diluted wines are good at iteration 9.

A t-test for the mean was carried out on the synthetic data for the attribute proline from iterations 4 and 9. From the test equal variance cannot be assumed, because F=8.004 with significance=0.004 for the data at iteration 4. Table 29 shows the result of the t-test for the oversampled data at iteration 4.

4.1.2.13 Statistical Test for Attribute Proline

Figure 13 shows the 5 box plots for the attribute 'proline'. We can see the final oversampling generated the data of a similar box. The oversampling at iteration 4 generated lower Q1, lower Q2, lower Q3, lower mean, and lower upper boundary.

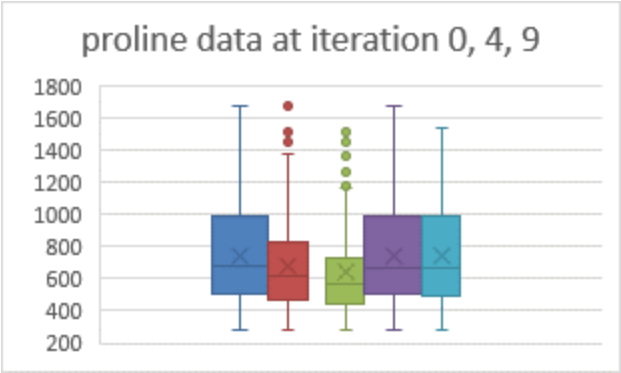

Fig. 13: Box plots for attribute 'proline' in the original data, the data of 10-fold CV and oversampled only at iteration 4, and the data of 10-fold CV and oversampled only at iteration 9 from left to right respectively

Table 29. Result of the t-test for the oversampled data at iteration 4 for the attribute proline

Data at	Mean	SD	t	p
Original	746.8933	314.9075	3.669	0.0
Iteration 4	643.5547	274.9652		

Because p=0.0, which is less than 0.05, we can see that the difference in the means is statistically significant, so we can say that the oversampled data only for attribute proline are not good at iteration 4. Because F=0.70 with significance=0.792 for the data at iteration 9, we can assume equal variance. Table 30 shows the result of the t-test for the oversampled data at iteration 9.

Table 30. Result of the t-test for the oversampled data at iteration 9 for the attribute proline

Data at	Mean	SD	t	p
Original	746.8933	314.9075	0.065	0.948
Iteration 9	745.2783	308.9427		

Because p=0.948, which is greater than 0.05, we can see that the difference in the means is statistically insignificant, so we can say that the oversampled data only for attribute proline are good at iteration 9.

Table 31 summarizes the t-test for the original data and oversampled data of 13 attributes at iteration 4 and 9.

Table 31. Summary of t-test for the original and oversampled data of 13 attributes of wine data

Attribute	Iteration 4		Iteration 9	
	t	p	t	p
Alcohol	4.084	**0.0**	-0.297	0.761
Malic acid	1.861	0.064	-1.051	0.294
ash	2.675	**0.008**	0.301	0.763
Alkalinity of ash	-1.095	0.275	-0.263	0.793
Magnesium	2.764	**0.006**	0.749	0.454
Total phenols	0.283	0.777	-0.069	0.945
Flavonoids	-0.147	0.883	0.102	0.919
Nonflavonoid phenols	-0.046	0.963	-0.113	0.910
Proantho-cyanins	0.123	0.902	-0.462	0.645
Color intensity	4.202	**0.0**	0.045	0.964
Hue	-1.830	0.068	0.334	0.739
OD280 or OD315 of diluted wines	-1.127	0.261	0.148	0.883
Proline	3.669	**0.0**	0.065	0.948

In summary, the oversampled data at iteration 9 are good for all 13 attributes of the wine data set, while only 8 attributes are good and 5 attributes are not good as indicated by p values at iteration 4 in the statistical sense. Moreover, all the t values of attributes are better at iteration 9 than iteration 4 except for attribute nonflavonoid phenols.

5 Conclusion

Building the most accurate machine learning model for a given dataset is a common goal of machine learning researchers, so several sites have provided experimental data for this purpose, and the UCI machine learning repository is one of them. The site has a variety of datasets and the size of the datasets range from small to very large. When we do not have sufficiently sized data, we may rely on oversampling. However, because simple oversampling increases the likelihood of overfitting by introducing replicated samples, several oversampling methods based on synthetic samples have been suggested, and SMOTE is one of the representative oversampling methods that can generate synthetic instances or samples of a minor class. Until now, oversampled data has been used conventionally to train machine learning models without statistical analysis, so it is not certain that the machine learning models will be fine for unseen cases in the future. However, because such synthetic data is different from the original data, we may wonder how much it resembles the original data so that the synthetic data may be used to improve machine learning models. In this sense, it is necessary to compare the synthetic data with the original data to see if it is statistically reliable. For this purpose, I conducted the study on a representative dataset called wine data in the UCI machine learning repository, which is one of the datasets that has been widely used by many researchers in research for various knowledge discovery models.

On the other hand, since training a machine learning model by supplying more high-quality training instances increases the probability of obtaining a machine learning model with higher accuracy, it was also checked whether a better machine learning model of random forests that is one of representative machine learning models can be obtained by generating much more synthetic data than the original data and using it for training the random forests. As summarized in Table 31, the results of the experiment showed that small-scale oversampling like oversampling at iteration 4 could produce synthetic data with statistical characteristics that were statistically slightly different from the original data, but when the oversampling rate was relatively high with the suggested method, it could be possible to generate synthetic data with statistical characteristics similar to the original data, and by using it to train the random forests, it could be possible to generate random forests with higher accuracy than using the original data alone. In this sense, this paper has contributed to the fact that it provides a methodology to increase the reliability of the machine learning models built using such oversampled data by analyzing the statistical characteristics of the oversampled data. So, we could supply statistically more reliable synthetic or oversampled data by applying the statistical methods. Moreover, the proposed method is applicable in any field where the conditional attributes are composed of numerical values and the results of the machine learning algorithm based on the original data are more or less good, so it can be applied in any field where oversampling is desired.

Our study can be applied to the oversampling method that generates synthetic data by the nearest neighbor information and the interpolation method when the attributes are numerical attributes, such as the SMOTE algorithm. Future research will be based on how to determine the nearest neighbor if the attributes are categorical, and how to do over-sampling based on that information. Moreover, if there is a large difference in the mean of attribute values for each class, the oversampled data generated by the proposed method may not satisfy the t-test, so in such a case, a balanced oversampling may be preferred.

Acknowledgment:
This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2023.

References:

[1] S. Tufail, H. Riggs, M. Tariq, A.I. Sarwat, Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms, *Electronics*, Vol. 12, Issue 8, 1789, 2023, pp. 1-43, https://doi.org/10.3390/electronics12081789.

[2] M. Kelly, R. Longjohn, K. Nottingham, *The UCI Machine Learning Repository*, https://archive.ics.uci.edu (Accessed Date: April 1, 2024).

[3] S. Aeberhard, M. Forina, Wine, *UCI Machine Learning Repository*, https://archive.ics.edu/dataset/109/wine, https://doi.org/10.24432/C5PC7J (Accessed Date: January 5, 2024).

[4] C. Sager, C. Janiesch, P. Zschech, A survey of image labelling for computer vision applications, *Journal of Business Analytics*, Vol. 4, No. 2, 2021, pp. 91-110, https://doi.org/10.1080/2573234X.2021.1908861.

[5] S. Imori, H. Shimodaira, An Information Criterion for Auxiliary Variable Selection in Incomplete Data Analysis, *Entropy*, Vol. 12, Issue 3, 281, 2019, pp. 1-19, https://doi.org/10.3390/e21030281.

[6] D.K. Jana, P. Bhunia, S.D. Adhikary, A. Mishra, Analyzing of salient features and classification of wine type based on quality through various neural network and support vector machine classifiers, *Results in Control and Optimization*, Vol. 11, 100219, 2023, pp. 1-33, https://doi.org/10.1016/j.rico.2023.100219.

[7] H. Li, X. Ye, A. Imakura, T. Sakuri, Ensemble Learning for Spectral Clustering, *2020 IEEE International Conference on Data Mining(ICDM)*, Sorrento, Italy, 17-20 November 2020, pp. 1094-1099, DOI: 10.1109/ICDM50108.2020.00131.

[8] X. Di, P. Yu, R. Bu, M. Sun, Mutual Information Maximization in Graph Neural Networks, 2020 *IEEE International Joint Conference on Neural Network(IJCNN)*, Glasgow, UK, 19-24 July 2020, pp. 1-17, DOI: 10.1109/IJCNN48605.2020.9207076.

[9] V. Ojha, G. Nicosia, Multi-objective Optimisation of Multi-output Neural Trees, *IEEE Congress on Evolutionary Computation*, Glasgow, Scotland(Online), 19-24 July 2020, pp. 1-8, https://doi.org/10.1109/CEC48606.2020.9185600.

[10] M. Lichouri, M. Abbas, Simple vs Oversampling-based Classification Methods for Fine Grained Arabic Dialect Identification in Twitter, *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, Barcelona, Spain (Online), December 2020, pp. 250-256.

[11] T. Wongvorachan, S. He, O. Bulut, A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining, *Information 2023*, Vol. 14, Issue 1, 54, 2023, pp. 1-15, https:///doi.org/10.3390/info14010054.

[12] N.V. Chawla, K.W. Dowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Synthetic Intelligence Research*, Vol. 16, 2002, pp. 321-357.

[13] H. Han, W. Wang, B. Mao, Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Sets Learning, In: Huang, DS., Zhang, XP., Huang, GB. (eds) *Advances in Intelligent Computing (ICIC 2005)*, 23-26 August 2005, Heifei, China, *Lecture Notes in Computer Science*, Vol. 3644, 2005, pp. 878-887, https://doi.org/10.1007/11538059_91.

[14] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *IEEE International Joint Conference on Neural Networks*, 1-6 June 2008, Hong Kong, China, 2008, pp. 1322-1328.

[15] Z. Zheng, Y. Cai, Y. Li, Oversampling method for imbalanced classification, *Computing and Informatics*, Vol. 34, 2015, pp. 1017-1037.

[16] L. Breiman, Random Forests, *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.

[17] A. Lulli, L. Oneto, D. Anguita, Mining Big Data with Random Forests, *Cognitive Computation,* Vol.11, 2019, pp. 294-316.

[18] R. Shiroyama, M. Wang, C. Yoshimura, Effect of sample size on habit suitability estimation using random forests: a case of bluegill, Lepomis macrochirus, *International Journal of Limnology*, Vol. 56, Article 13, 2020, https://doi.org/10.1051/limn/2020010.

[19] C. Chi, P. Vossler, Y. Fan, J. Lv, Asymptotic Properties of High-Dimensional Random

Forests, *The Annals of Statistics*, Vol. 50, No. 6, 2022, pp. 3415-3238, DOI: 10.1214/22-AOS2234.

[20] How to Run Levene's Test in SPSS? https://www.spss-tutorials.com/levenes-test-in-spss/ (Accessed Date: February 1, 2024).

[21] A. Field, *Discovering Statistics Using IBM SPSS Statistics: North American Edition*, 5th ed., SAGE Publications Ltd., 2017.

[22] Independent t-test using SPSS Statistics, https://statistics.laerd.com/spss-tutorials/independent-t-test-using-spss-statistics.php (Accessed Date: February 1, 2024).

[23] E. Frank, M.A. Hall, I.H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Fourth Edition, 2016.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The sole author contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2023.

Conflict of Interest

The author has no conflicts of interest to declare.