Ozone Day Classification using Random Forests with Oversampling and Statistical Tests

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

Abstract: - Accurate warning of ozone concentration levels in the air is very important for public health. However, the characteristics of the public data related to ozone level detection in the UCI machine learning repository make it difficult to build a warning system based on machine learning techniques. The data consists of 72 relatively large numerical attributes and are measured and collected for 7 years with some blank data, and the distribution of ozone days and normal days is very unbalanced, making it difficult to create an accurate classification model. In this paper to solve the high dimensional attribute problem PCA is applied first, resulting in the 72 attributes being reduced to 20 attributes, and generating slightly better random forests, but the classification for ozone days is still poor due to insufficient data. To solve the insufficient data problem for the minor class which is 6.3% of the total, SMOTE which is one of the representative oversampling methods is applied to a minor class at very high rates repeatedly. It was also checked whether a better machine learning model of random forests can be obtained after applying oversampling at the same very high rate for each class, generating much more synthetic data than the original data and using it to train the random forests. In addition, to ensure the reliability of the synthetic data generated by SMOTE statistical test has been done for each attribute to see if it is statistically reliable. The results of the experiment showed that when the oversampling rate was relatively high with the suggested oversampling and statistical tests, 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 and more balanced classification accuracy than using the original data alone, from 94% to 100%. In this sense, this paper has contributed that it provides a methodology to increase the reliability of the machine learning model of random forests for very skewed and high dimensional data like the ozone day classification dataset.

Key-Words: - Ozone level detection, numerical attributes, high dimensionality of data, PCA, data skewness, oversampling, box plot, t-test, random forests.

Received: June 26, 2024. Revised: November 11, 2024. Accepted: December 10, 2024. Published: December 30, 2024.

1 Introduction

At the top 20 kilometers of the Earth's atmosphere, the ozone layer surrounds the Earth. The ozone layer absorbs ultraviolet rays from the sun and prevents adverse effects on life on Earth, so it is essential for the survival of life. However, ozone produced near the earth's surface is a factor that worsens our health. Nitrogen oxides and hydrocarbons emitted from factories and automobile exhaust gases can be decomposed by strong sunlight, and ozone can be produced when the oxygen molecules are combined with oxygen in the air. Ozone, which is produced near the earth's surface, can irritate the respiratory organs, irritate the eyes, and pose a health risk to respiratory patients. Normally, the ozone concentration on a normal day is maintained at the level of 0.02 ppm, but if the ozone concentration rises and there is more than a certain level of ozone in the atmosphere, ozone action alerts are issued according to the ozone concentration. Meanwhile, discovering knowledge models of highly accurate ozone level alarm forecasting models for the Houston area using a dataset called ‘ozone level detection dataset’ has attracted researchers’ attention where several technical difficulties exist like high dimensionality and data skewness, and also some blank data, [1].

The first difficulty in discovering good knowledge models is that ozone days are only 2% or 6% level in the dataset among around 2500 data entries that were collected over seven years. In other words, the data has a very skewed class distribution, which usually happens in real-world big data. For example, a lot of data is produced in the

semiconductor production process, but most of the processes produce good products rather than defective products, so there is a severe class imbalance in the data. The second difficulty is that the dataset might have multicollinearity between attributes and contain several irrelevant features among the 72 conditional features in the original dataset, and there are some missing data also which is common in real-world data. Therefore, finding highly accurate knowledge models for the dataset is a challenging task, and we may expect that the dataset may offer a chance to explore new data mining techniques, and to provide guidance for similar problems.

There are several ways to deal with the curse of dimensionality, [2], and Principle Component Analysis(PCA) is one of the most widely used methods for numerical attributes to deal with the problem, [3]. Therefore, it is necessary to test the applicability of PCA to the high-dimensional dataset such as the ozone dataset.

In addition, in classification problems, when the number of data instances markedly differs by class, it is common to apply oversampling to ensure that minority classes are not discriminated against classification. Meanwhile, it is generally true that the larger the high-quality data used for training, the higher the accuracy of knowledge models by machine learning algorithms can be. Considering that oversampling can generate more synthetic data than the original data, we want to see if feeding a lot of oversampled data into the training of a machine learning model could lead to a better machine learning model on the condition that statistical analysis will be performed to see how statistically reliable such oversampled data is, especially for the ozone dataset.

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

Ozone day prediction analysis in the original paper was performed with a precision-recall curve only by changing a subjective decision threshold of ozone day with eight different machine learning models--mostly decision tree-based models like C4.5, baggingC4.5, random decision forest, [1]. Most of the precision-recall ranges 0.4 ~ 0.6 in the experiment. Note that precision = TP/(TP+FP), and recall = TP/(TP+FN), where TP stands for the number of True Positives, FP stands for the number of False Positives, and FN stands for the number of False Negatives, and positive means an ozone day which is a rare case in the ozone dataset. Some other papers used different approaches to solve the problem. To cope with the high dimensionality of the data and missing data problem, the MGHFAmiss (Mixture of Generalized Hyperbolic Factor Analysis) model was invented and achieved the accuracy of 54.6% ~ 73.2% depending on the latent factors between 1 ~ 60 for the ozone dataset, [4]. In [5] various algorithms like Support Vector Machines, K Nearest Neighbours, XGBoost, LGBM, Hist Gradient Boosting Machine, and Deep Neural Networks were applied, and Extreme Gradient Boosting (XGBoost) algorithm showed the best results in accuracy of 95% for the dataset. Downward-growing neural network which can adapt its neural structure during training achieved an accuracy of 97.32% for the dataset, [6]. Note that ozone day cases occupy only 2.9% ~ 6.3% level in the whole cases, so just classifying no ozone day may satisfy around 94% ~ 97% accuracy. So, we still need more accurate classifiers, especially for the ozone day.

Next, let's take a look at the principle of the important techniques that will be used in this paper. Principal component analysis(PCA) is a technique that converts high-dimensional data into low-dimensional data and uses orthogonal transformations to convert data from high-dimensional spaces that are likely to be related to each other into the data of low-dimensional spaces that are not linearly related. When the data is mapped as a single axis, the axis with the largest variance is placed as the first main component, and the axis with the second largest variance is placed as the second main component, etc. In this way, the data are broken down into components that best represent the differences, [7].

The goal of a decision tree is to ensure that each terminal node consists of data instances of the same class as much as possible based on appropriate branching criteria for subtrees, and it is pruned at an appropriate level to avoid overfitting and reduce the size of the tree, [8]. The decision tree has the advantage of being relatively easy to understand because the machine learning results are expressed in a single tree structure, but since it is based on a greedy algorithm in generating subtrees, there is a possibility of missing the global optimal solution and overfitting may happen. Random forests were developed to improve the fact that even if you sacrifice good understandability, which is the advantage of decision trees, you can miss the global solution, which is a disadvantage of decision trees. Random forests generate many random decision

trees based on random sampling with replacement and use the many decision trees to classify by majority voting, [9]. If the sample size n is very large in sampling with replacement, it is known that the proportion of samples that are never sampled out of n samples is about 36.8%. In other words, when each decision tree is generated, some data is sampled several times and some data is not sampled at all, so only about 63.2% of the total data is randomly selected to generate the decision tree. As a result, as each decision tree is generated slightly differently for the overall data, it ends up with a slightly biased decision tree, so that it can be a knowledge model that represents the data from a slightly different perspective, making the random forests a set of knowledge models that better responds to the enormous size of the data space.

When each random decision tree is generated, a random selection of root attributes among a set of attributes that are selected randomly is performed for each subtree, and no pruning is performed. The default value for the size of the set is INT($\log_2$(the number of attributes) + 1) or one can give an appropriate size of the set. By randomly selecting the root attribute of each subtree, we may avoid local optimization due to the greedy search. Random forests are known to be one of the most reasonable machine-learning algorithms across a wide range of data, [10]. Other factors that affect the performance of random forests are the size and dimensionality of the data. Because a smaller sample size may not represent the population well, the sample size may affect the performance of random forests, [11], and the dimensionality of the data affects the size of the sample required to generate good random forests, [12].

Random forests are likely to be more accurate with a larger number of samples, so when a given number of data is fixed and not large, oversampling can be a means of obtaining a large amount of data. Simple oversampling oversamples a specific class to give more attention to training a machine learning model, [13]. Because simple oversampling increases the likelihood of overfitting by introducing replicated samples, oversampling based on synthetic samples was invented. SMOTE(Synthetic Minority Over-sampling Technique) is one of the representative oversampling methods that generate synthetic instances of a minor class. SMOTE selects k-nearest neighbors where k can be given by users, for example, k=5, and generates a synthetic data instance of interpolation by multiplying an interval value of continuous attributes of the k-nearest neighbors with a random number between 0 and 1, [14].

3 Problem Formulation

The ozone level detection dataset has 73 conditional attributes and one decisional attribute called class, and the first conditional attribute contains measured date information. The other 72 attributes are continuous or numerical conditional attributes. One decisional attribute classifies ozone level into two classes, 0 for normal day and 1 for ozone day. The dataset contains two datasets—eight-hour and one-hour forecast datasets. The class distribution is very skewed. For example, only 6.3% of data belongs to Ozone Day in the 8-hour ozone dataset, and only 2.9% of data belongs to Ozone Day in the 1-hour ozone data. The 8-hour ozone is the average of the ozone concentration of the past 8 hours, and vice versa. The data set originated from various meteorology and ozone data for the Houston, Galveston, and Brazoria areas. The 72 data attributes are extracted from several databases within two major federal data warehouses and one local database for air quality control. The dataset has been open to the public in the UCI machine learning repository since 2008, [15].

To build random forests for the dataset, a pre-processing like PCA will be performed first since the ozone dataset has a rather large number of attributes. Moreover, since the ozone data has a very skewed distribution for the ozone day class, we want to apply SMOTE to generate new synthetic data, 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 of random forests could 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 perform 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, [16]. 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, [17], 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, [18].

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 PCA, oversampling, and to generate random forests an open-source tool called Weka will be used. [19]. When principal components analysis and transformation of the data are performed, a ranker search based on eigenvalues is adapted in Weka. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance (default = 0.95) in the original data.

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 ozone level detection dataset – especially for the 8-hour ozone dataset because of the limitation of space. Note that the distribution of ozone days and normal days is very unbalanced, making it difficult to create an accurate classification model. To solve the insufficient data problem for the minor class which is 6.3% of the total, a progressive oversampling process will be performed, and PCA will be applied to reduce the number of attributes to a small number for ease of statistical analysis. The experimental procedure consists of three steps;

First, PCA will be applied to reduce the number of attributes.

Second, we repeat oversampling until the random forests are not much improved.

Finally, for each attribute, we'll draw box plots for several steps of oversampling, and statistical testing will be performed to see if there are any differences between the oversampled data and the original data.

Oversampling of 100% on the classes with the highest number of misclassifications in the resulting confusion matrix will be repeated until there is not much 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.

4 Experimentation

The 8-hour ozone dataset in the 'Ozone Level Detection' dataset in the UCI machine learning repository, [15], is used for experiments. The goal of this experiment is to find the best models of random forests as accurately as possible based on oversampling and to perform statistical tests on the oversampled data.

4.1 Ozone level Detection Dataset

The data set is the collection of various meteorology and ozone data for the Houston, Galveston, and Brazoria areas, [1]. Because of the limitations of space, among the eight-hour and one-hour ozone data, the eight-hour ozone data will be analyzed. The 8-hour ozone is the average of the ozone concentration of the past 8 hours. Since the two data sets have the same characteristics, the one-hour ozone dataset could be analyzed similarly. The eight-hour data set has 2,534 records and has 73 conditional attributes and one decisional attribute, named class. Among the 73 attributes, the first attribute represents a date, so it will be ignored for the analysis. Among 2,534 records 2374 records represent normal day (93.7%) and 160 records represent ozone day (6.3%), so the data distribution is very skewed. The 72 conditional attributes consist of numerical attributes as in Table 1 (Appendix).

In the first row of Table 1 (Appendix), 'D. values' means distinct values, and 'SD' means Standard Deviation. The 72 attributes contain various measures of air pollutants and meteorological information for the area. From the top of Table 1 (Appendix), 11% ~ 12% of the WSR series have missing values. 7% ~ 8% of T series, T0 ~ T_AV, have missing values. 4% ~ 8% of T85 ~ SLP_ have missing values. Attribute Precp has no missing values.

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 ozone data set

Accuracy in 10-fold CV (%)	Confusion matrix		No. of Misclassified
93.6464	2373	1	161
	160	0	

Note that 2374 records represent no ozone day and 160 records represent ozone day, so the trained model using the original dataset cannot classify ozone day at all. It just classifies 'no ozone day' with an accuracy of 93.6464%.

4.1.1 Principle Component Analysis

Because the original data set contains many similar attributes, principle component analysis is performed. The principle component analysis module in Weka is used with an attribute ranker algorithm and dimensionality reduction is done by choosing enough eigenvectors that can account for 95% (default) of the variance in the original data. Table 3 shows the result of PCA.

Table 3. The result of PCA for the original data of the ozone data set

Rank	Attribute
1	-0.179T_AV-0.176T9-0.176T10- ...
2	0.243WSR_AV+0.209WSR_PK +0.184WSR10+ …
3	0.247KI+0.243RH85+0.24 RH70+ …
4	0.231HT85+0.217WSR8+0.21WSR7+ …
5	0.376HT85+0.31 HT70+0.293SLP- …
6	-0.264WSR14-0.258WSR13-0.24RH70- …
7	0.443V70+0.38 V85+0.367V50- …
8	-0.397U85-0.356SLP-0.354U70- …
9	0.287WSR9+0.267WSR8-0.259WSR0- …
10	-0.558SLP_-0.296T50-0.29T70+ …
11	0.729Precp+0.421SLP_-0.226RH85+ …
12	-0.332RH50-0.315V50+0.3 Precp+ …
13	-0.491RH50+0.236RH85+0.223WSR18- …
14	-0.534SLP_+0.35 U85-0.274V50+ …
15	-0.386RH50+0.251Precp+0.247RH70- …
16	0.373WSR6+0.344WSR5-0.336WSR0- …
17	0.536RH70-0.442RH50-0.328TT- ..
18	0.334WSR19-0.311WSR23-0.265WSR16- ..
19	0.479U50+0.372RH85+0.325T50- …
20	-0.293V50-0.281U50+0.277RH85+ …

The above attributes are the abbreviations of the original expressions generated from Weka. The original expressions use all the 72 original attributes but for the sake of brevity, the first three terms in the original expression were cited in the table.

The attributes in the data set have different mean and standard deviation(SD) for each class.

Table 4 summarizes the mean and SD for each class. The name of attributes is numbered according to their rank in Table 3 for notational convenience in Table 4.

Table 4. Mean and SD of each attribute for each class

Att.	Class value 0		Class value 1	
	mean	SD	mean	SD
1	0.311254	5.468539	-4.61823	3.257799
2	0.140408	3.868573	-2.08331	2.288955
3	0.04811	2.345631	-0.71383	1.668737
4	-0.00913	1.825662	0.135536	1.327391
5	0.021092	1.669574	-0.31295	1.363076
6	-0.0258	1.406649	0.382803	1.249007
7	0.022331	1.275332	-0.33133	1.206614
8	-0.04407	1.067737	0.653928	1.030845
9	0.012374	1.055998	-0.18359	0.839559
10	0.007585	0.989363	-0.11255	0.735529
11	0.001341	0.978297	-0.01989	0.540886
12	0.010842	0.901597	-0.16087	0.786119
13	-0.01166	0.872	0.173068	0.712733
14	-0.01043	0.835519	0.0154747	0.718726
15	-0.01338	0.761486	0.198496	0.598059
16	-0.00783	0.756101	0.116105	0.538627
17	0.003712	0.694192	-0.05508	0.673037
18	0.004504	0.619874	-0.06682	0.564633
19	-0.00015	0.584919	0.002162	0.527732
20	0.020751	0.054201	-0.3079	0.520444

Figure 1 shows the box plot of attribute values for class 0. In the figure the order of each attribute is from left to right: attribute 1, attribute 2, ..., and attribute 20.

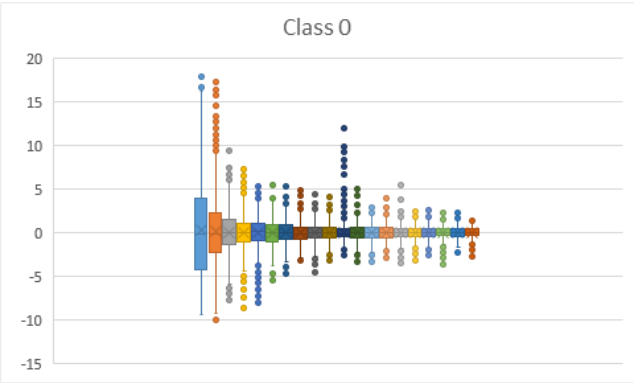

Fig. 1: The box plot of attribute values for class 0.

Figure 2 shows the box plot of attribute values for class 1.

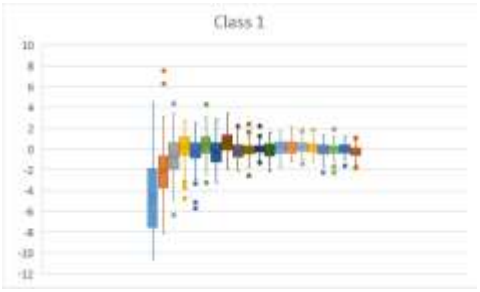

Fig. 2: The box plot of attribute values for class 0

As we can see in Figure 1 and Figure 2, it can be seen that there is a difference in value for each class depending on the rank of the attribute by PCA.

Table 5 shows random forests for the transformed data of ozone data by PCA.

Table 5. The result of random forests for the transformed data of the ozone data set by PCA

Accuracy in 10-fold CV (%)	Confusion matrix		No. of Misclassified
94.041	2370	4	151
	147	13	

The transformed data shows slightly better random forests than the original data as we compare it with Table 2. But, ozone day classification is still very poor.

4.1.2 Progressive and Repetitive Oversampling

The target data of oversampling is the data generated by the PCA. The following is the progressive and repetitive oversampling procedure.

Procedure:

1. Pick a class for oversampling with the highest number of misclassifications for the next oversampling in the random forests of the original data;
2. Perform the oversampling of 800%;
   A. Generate the random forests using the original and oversampled data together and test based on 10-CV;
   B. Generate the random forests using the oversampled data alone and test based on the original data;
3. **Repeat**
   A. Pick a class for oversampling with the highest number of misclassifications for the next oversampling in the 10-CV of the last random forests;
   B. Perform the oversampling of 100%;
      i. Generate the random forests using the original and oversampled data together and test based on 10-CV;
      ii. Generate the random forests using the oversampled data alone and test based on the original data;
4. **Until** the random forests are not much improved;

Because the data is very skewed, initially oversampling of 800% by SMOTE is applied for the minority which has a class value of 1, and after that oversampling of 100% for the classes with the highest number of misclassifications in the confusion matrix is repeated until random forests are not much improved. 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(10-CV) is considered when original data and oversampled data are used together for training and testing, and the number of misclassifications of the random forests that are trained by oversampled instances only and tested by the original data is considered also. 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 may pick one of them randomly. Table 6 and Table 7 show the results of the experiment, and Table 6 shows the result of the experiment using 10-fold cross-validation.

Table 6. The result of progressive oversampling with 10-fold CV for ozone data

Iteration No.	Oversampled classes	Accuracy in 10-CV (%)	Confusion matrix		No. of Misclassified
1	2	94.6513	2250	124	204
			80	1360	
2	1	96.6387	4670	78	208
			130	1310	
3	2	97.8894	4634	114	151
			47	2833	
4	1	98.9415	9443	53	131
			78	2802	
5	2	99.3249	9428	68	103
			35	5725	
6	1	99.6849	18965	27	78
			51	5709	
7	2	99.8329	18961	31	51
			20	11500	
8	1	99.9333	37978	6	33
			27	11493	
9	2	99.9705	37976	8	18
			10	23030	

Table 7 shows the corresponding result of the experiment when we use over-sampled instances only for training and the original data for testing.

Table 7. The corresponding result of progressive oversampling when oversampled data are used for training and the original data are used for testing for the ozone data set

Itera-tion No.	Over sampled classes	Accuracy in 10-CV (%)	Confusion matrix		No. of Misclassified
1	2	6.3141	0	2374	2374
			0	160	
2	1	97.0797	2312	62	74
			12	148	
3	2	96.1721	2277	97	97
			0	160	
4	1	98.895	2351	23	28
			5	155	
5	2	99.0529	2352	22	24
			2	158	
6	1	99.4081	2363	11	15
			4	156	
7	2	99.487	2363	11	13
			2	58	
8	1	99.7632	2370	4	6
			2	158	
9	2	99.7238	2368	6	7
			1	159	

Note that iterations 8 and 9 generated almost similar results in Table 7. We may select the random forests at iteration 8 because the number of misclassified cases is smaller in the test done by the original data.

Because the attribute values are different by class as we can see in Figure 1 and Figure 2, the equal oversampling rates for each class of 800% and 7200% which are the oversampling rate for class 1 at iteration 1 and iteration 8 have been performed before we perform statistical tests. Table 8, Table 9, Table 10 and Table 11 show the results. Table 8 shows the random forests with an equal oversampling rate of 800% in 10-fold cross-validation.

Table 8. The result of random forests of the ozone data set with the equal oversampling rate of 800% for each class

Accuracy in 10-fold CV (%)	Confusion matrix		No. of Misclassified
99.3425	21354	10	378
	368	1072	

Note that the TP rate for ozone day is 0.744, while the TP rate for no ozone day is 1.0. Table 9 shows the random forests when oversampled data only is used for training and the original data is used for testing with an equal oversampling rate of 800% for each class.

Table 9. The result of random forests of the ozone data set with the equal oversampling rate of 800% for each class when oversampled data only is used for training

Accuracy when the original data is used for testing(%)	Confusion matrix		No. of Misclassified
99.3425	2374	0	42
	42	118	

Note that the TP(True Positive) rate for ozone day is 0.738, while the TP rate for no ozone day is 1.0.

Table 10 shows the random forests with an equal oversampling rate of 7200% in 10-fold cross-validation.

Table 10. The result of random forests of the ozone data set with the equal oversampling rate of 7200% for each class

Accuracy in 10-fold CV (%)	Confusion matrix		No. of Misclassified
99.8113	173298	4	349
	345	11335	

Note that the TP rate for ozone day is 0.97, while the TP rate for no ozone day is 1.0.

Table 11 shows the random forests when oversampled data only is used for training and the original data is used for testing.

Table 11. The result of random forests of the ozone data set with the equal oversampling rate of 7200% for each class when oversampled data only is used for training

Accuracy when the original data is used for testing(%)	Confusion matrix		No. of Misclassified
99.8421	2374	0	4
	4	156	

Note that the TP rate for ozone day is 0.975, while the TP rate for no ozone day is 1.0.

4.1.3 Statistical Test for Oversampled Data of Ozone

Statistical tests are done to see the properties of oversampled data for each attribute. Equally oversampled data of the oversampling rate of 800% and 7200% are used for box plot, and 7200% are used for t-test.

4.1.3.1 Statistical Test for the 1st Attribute

Figure 3 shows the 5 box plots for the 1st attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively. Note that the data of the 10-fold CV contains the original data as well as oversampled data. We can see the oversampling generated similar boxes.

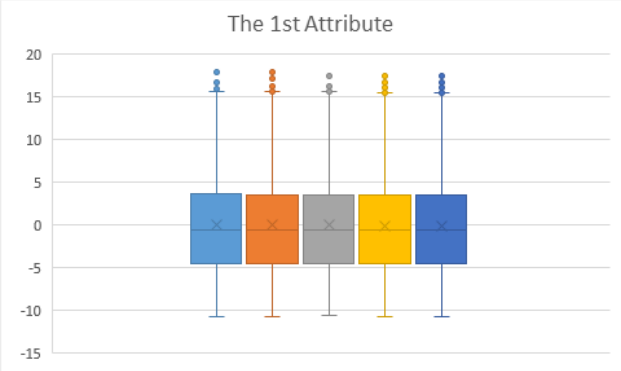

Fig. 3: Box plots for the 1st attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 1st attribute at the oversampling rate of 7200%. Note that the data contains the original data and synthetic data together. In the box plot in Figure 3, it's the 4th one. From the test equal variance was assumed, because F=0.912 with significance=0.34 for the data. Table 12 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%. SD means standard deviation in the table. The original numerical value that was not omitted was -0.000000003946330 for the mean of the original data, so for the sake of simplicity of expression, only 4 decimal places after the point are written.

Table 12. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 1st attribute

Data at	Mean	SD	t	p
Original	-0.0000	5.4882	0.831	0.406
Over-sampled	-0.0901	5.4197		

Because p=0.406 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 the 1st attribute are good at the oversampling rate of 7200%.

4.1.3.2 Statistical Test for the 2nd Attribute

Figure 4 shows the 5 box plots for the 2nd attribute. We can see the oversampling generated slightly narrow boxes.

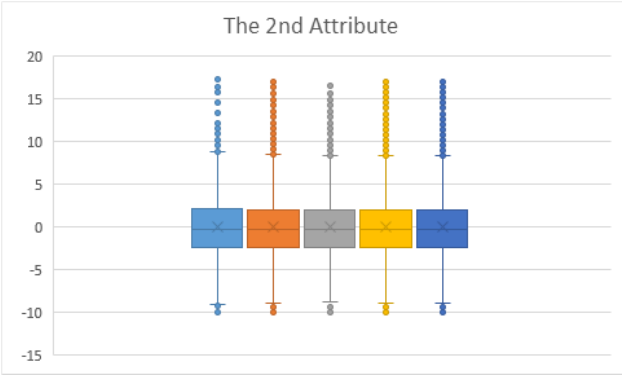

Fig. 4: Box plots for the 2nd attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 2nd attribute at the oversampling rate of 7200%. From the test equal variance was assumed, because F=3.706 with significance=0.054 for the data. Table 13 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 13. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 2nd attribute

Data at	Mean	SD	t	p
Original	-0.0000	3.8265	0.891	0.373
Over-sampled	-0.0661	3.7097		

Because p=0.373 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 the 2nd attribute are good at the oversampling rate of 7200%.

4.1.3.3 Statistical Test for the 3rd Attribute

Figure 5 shows the 5 box plots for the 3rd attribute. We can see the oversampling generated the data of a slightly narrow distribution.

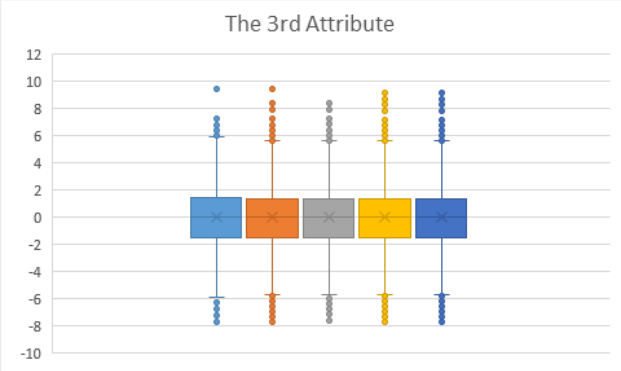

Fig. 5: Box plots for the 3rd attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 3rd attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=10.222 with significance=0.001 for the data. Table 14 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 14. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 3rd attribute

Data at	Mean	SD	t	p
Original	-0.0000	2.3159	0.900	0.368
Over-sampled	-0.0417	2.1955		

Because p=0.368 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 the 3rd attribute are good at the oversampling rate of 7200%.

4.1.3.4 Statistical Test for the 4th Attribute

Figure 6 shows the 5 box plots for the 4th attribute. We can see the oversampling generated the data of a slightly narrow distribution.

Fig. 6: Box plots for the 4th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 4th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=33.769 with significance=0.000 for the data. Table 15 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 15. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 4th attribute

Data at	Mean	SD	t	p
Original	-0.0000	1.7984	-0.495	0.621
Over-sampled	-0.0178	1.6325		

Because p=0.621 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 the 4th attribute are good at the oversampling rate of 7200%.

4.1.3.5 Statistical Test for the 5th Attribute

Figure 7 shows the 5 box plots for the 5th attribute. We can see the oversampling generated slightly narrow boxes.

Fig. 7: Box plots for the 5th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 5th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=49.555 with significance=0.000 for the data. Table 16 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 16. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 5th attribute

Data at	Mean	SD	t	p
Original	0.0000	1.7984	-0.864	0.388
Over-sampled	0.02855	1.6325		

Because p=0.388 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 the 5th attribute is good at the oversampling rate of 7200%.

4.1.3.6 Statistical Test for the 6th Attribute

Figure 8 shows the 5 box plots for the 6th attribute. We can see the oversampling generated slightly narrow boxes.

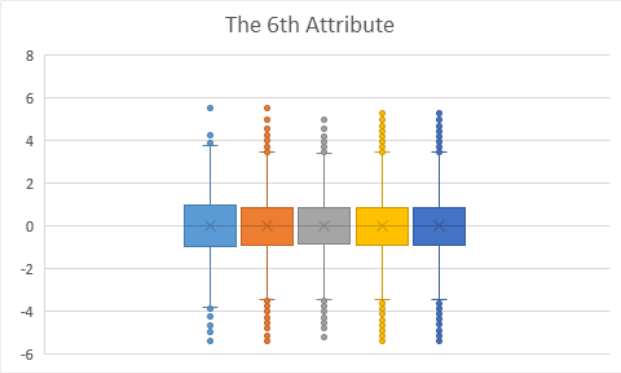

Fig. 8: Box plots for the 6th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 6th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=55.760 with significance=0.000 for the data. Table 17 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 17. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 6th attribute

Data at	Mean	SD	t	p
Original	-0.0000	1.4005	0.368	0.713
Over-sampled	-0.0103	1.2536		

Because p=0.713 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 the 6th attribute are good at the oversampling rate of 7200%.

4.1.3.7 Statistical Test for the 7th Attribute

Figure 9 shows the 5 box plots for the 7th attribute. We can see the oversampling generated slightly narrow boxes.

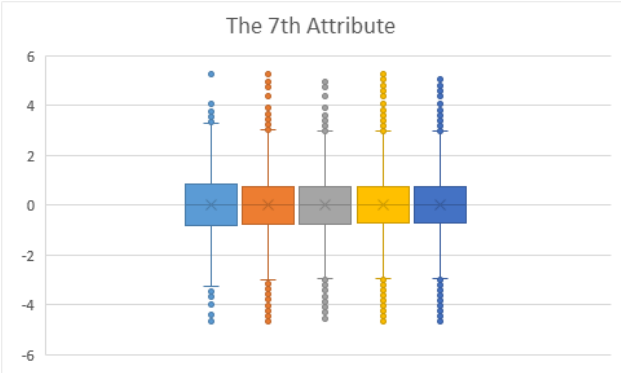

Fig. 9: Box plots for the 7th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 7th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=62.711 with significance=0.000 for the data. Table 18 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 18. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 7th attribute

Data at	Mean	SD	t	p
Original	-0.0000	1.2738	-0.679	0.498
Over-sampled	-0.0173	1.1324		

Because p=0.498 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 the 7th attribute are good at the oversampling rate of 7200%.

4.1.3.8 Statistical Test for the 8th Attribute

Figure 10 shows the 5 box plots for the 8th attribute. We can see the oversampling generated slightly narrow boxes.

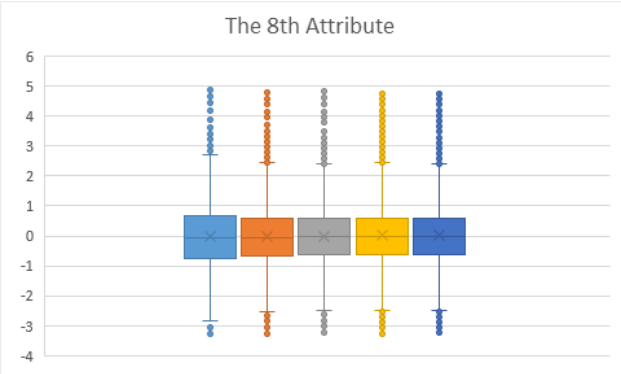

Fig. 10: Box plots for the 8th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 8th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=91.051 with significance=0.000 for the data. Table 19 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 19. the result of the t-test for the oversampled data at the oversampling rate of 7200% for the 8th attribute

Data at	Mean	SD	t	p
Original	0.0000	1.0787	-0.059	0.953
Over-sampled	0.0013	0.9357		

Because p=0.953 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 the 8th attribute are good at the oversampling rate of 7200%.

4.1.3.9 Statistical Test for the 9th Attribute

Figure 11 shows the 5 box plots for the 9th attribute. We can see the oversampling generated slightly narrow boxes.

Fig. 11: Box plots for the 9th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 9th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=105.977 with significance=0.000 for the data. Table 20 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 20. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 9th attribute

Data at	Mean	SD	t	p
Original	0.0000	1.0787	0.861	0.389
Over-sampled	-0.0180	0.8922		

Because p=0.389 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 the 9th attribute are good at the oversampling rate of 7200%.

4.1.3.10 Statistical Test for the 10th Attribute

Figure 12 shows the 5 box plots for the 10th attribute. We can see the oversampling generated the slightly narrow boxes.

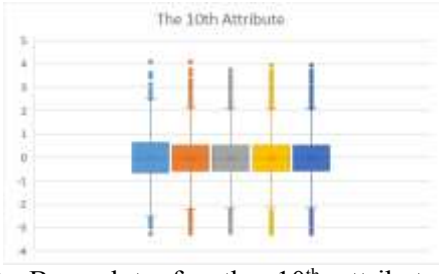

Fig. 12: Box plots for the 10th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 10th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=86.864 with significance=0.000 for the data. Table 21 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 21. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 10th attribute

Data at	Mean	SD	t	p
Original	-0.0000	0.9756	0.889	0.374
Over-sampled	-0.0173	0.8481		

Because p=0.374 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 the 10th attribute are good at the oversampling rate of 7200%.

4.1.3.11 Statistical Test for the 11th Attribute

Figure 13 shows the 5 box plots for the 11th attribute. We can see the data range is narrower than other attributes, and the oversampling generated slightly narrow boxes.

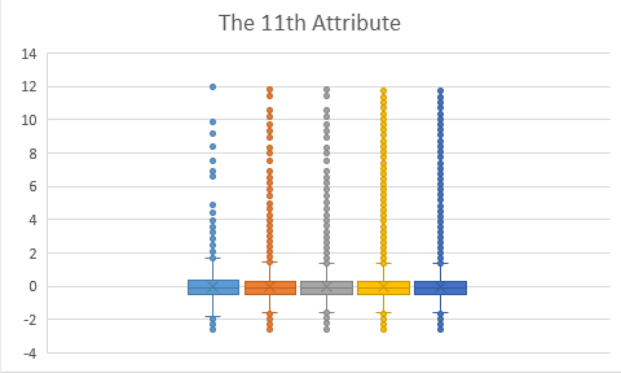

Fig. 13: Box plots for the 11th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 11th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=48.838 with significance=0.000 for the data. Table 22 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 22. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 11th attribute

Data at	Mean	SD	t	p
Original	-0.0000	0.9566	2.138	0.033
Over-sampled	-0.0408	0.8353		

Because p=0.033 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 for the 11th attribute are not good at the oversampling rate of 7200%. From this, we can see that the range of quartiles containing most of the data is narrow, so the influence of outliers on oversampling played a large role.

4.1.3.12 Statistical Test for the 12th Attribute

Figure 14 shows the 5 box plots for the 12th attribute. We can see the oversampling generated slightly narrow boxes.

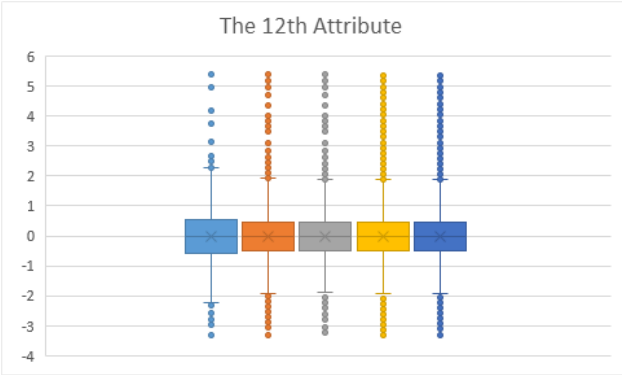

Fig. 14: Box plots for the 12th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 12th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=125.825 with significance=0.000 for the data. Table 23 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 23. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 12th attribute

Data at	Mean	SD	t	p
Original	0.0000	0.8956	0.579	0.563
Over-sampled	-0.0103	0.7597		

Because p=0.563 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 the 12th attribute are good at the oversampling rate of 7200%.

4.1.3.13 Statistical Test for the 13th Attribute

Figure 15 shows the 5 box plots for the 13th attribute. We can see the oversampling generated slightly narrow boxes.

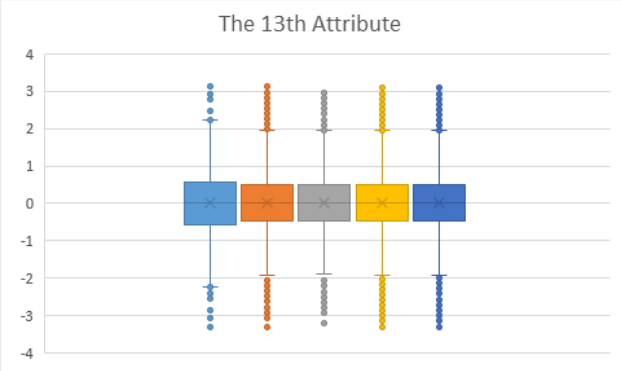

Fig. 15: Box plots for the 13th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 13th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=130.250 with significance=0.000 for the data. Table 24 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 24. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 13th attribute

Data at	Mean	SD	t	p
Original	0.0000	0.8639	-1.222	0.222
Over-sampled	0.0211	0.7317		

Because p=0.222 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 the 13th attribute are good at the oversampling rate of 7200%.

4.1.3.14 Statistical Test for the 14th Attribute

Figure 16 shows the 5 box plots for the 14th attribute. We can see the oversampling generated slightly narrow boxes.

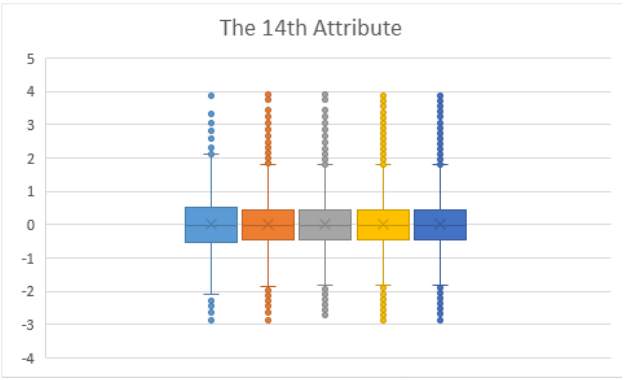

Fig. 16: Box plots for the 14th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 14th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=120.384 with significance=0.000 for the data. Table 25 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 25. the result of the t-test for the oversampled data at the oversampling rate of 7200% for the 14th attribute

Data at	Mean	SD	t	p
Original	0.0000	0.8639	0.449	0.653
Over-sampled	-0.0074	0.7317		

Because p=0.653 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 the 14th attribute are good at the oversampling rate of 7200%.

4.1.3.15 Statistical Test for the 15th Attribute

Figure 17 shows the 5 box plots for the 15th attribute. We can see the oversampling generated slightly narrow boxes.

Fig. 17: Box plots for the 15th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 15th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=86.782 with significance=0.000 for the data. Table 26 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 26. the result of the t-test for the oversampled data at the oversampling rate of 7200% for the 15th attribute

Data at	Mean	SD	t	p
Original	0.0000	0.8639	-0.298	0.766
Over-sampled	-0.0074	0.7317		

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 the 15th attribute are good at the oversampling rate of 7200%.

4.1.3.16 Statistical Test for the 16th Attribute

Figure 18 shows the 5 box plots for the 16th attribute. We can see the oversampling generated slightly narrow boxes.

Fig. 18: Box plots for the 16th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 16th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=137.999 with significance=0.000 for the data. Table 27 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 27. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 16th attribute

Data at	Mean	SD	t	p
Original	0.0000	0.7448	0.242	0.809
Over-sampled	-0.0036	0.6193		

Because p=0.809 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 the 16th attribute are good at the oversampling rate of 7200%.

4.1.3.17 Statistical Test for the 17th Attribute

Figure 19 shows the 5 box plots for the 17th attribute. We can see the oversampling generated slightly narrow boxes.

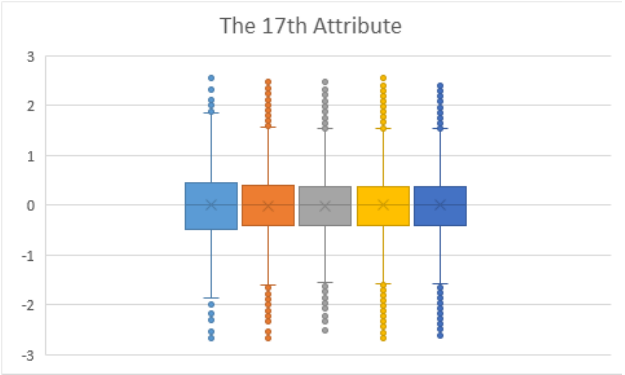

Fig. 19: Box plots for the 17th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 17th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=134.689 with significance=0.000 for the data. Table 28 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 28. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 17th attribute

Data at	Mean	SD	t	p
Original	-0.0000	0.7448	0.151	0.880
Over-sampled	-0.0021	0.6193		

Because p=0.880 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 the 17th attribute are good at the oversampling rate of 7200%.

4.1.3.18 Statistical Test for the 18th Attribute

Figure 20 shows the 5 box plots for the 18th attribute. We can see the oversampling generated slightly narrow boxes.

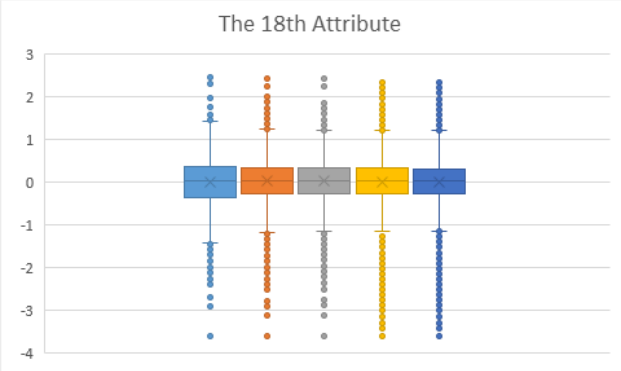

Fig. 20: Box plots for the 18th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 18th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=181.415 with significance=0.000 for the data. Table 29 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 29. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 18th attribute

Data at	Mean	SD	t	p
Original	0.0000	0.6167	-0.944	0.345
Over-sampled	0.0116	0.4983		

Because p=0.345 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 the 18th attribute are good at the oversampling rate of 7200%.

4.1.3.19 Statistical Test for the 19th Attribute

Figure 21 shows the 5 box plots for the 19th attribute. We can see the oversampling generated slightly narrow boxes.

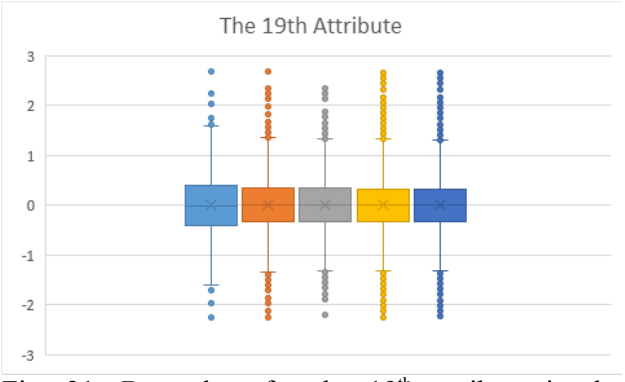

Fig. 21: Box plots for the 19th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 19th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=169.748 with significance=0.000 for the data. Table 30 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 30. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 19th attribute

Data at	Mean	SD	t	p
Original	-0.0000	0.5814	-0.135	0.893
Over-sampled	0.0016	0.4880		

Because p=0.893 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 the 19th attribute are good at the oversampling rate of 7200%.

4.1.3.20 Statistical Test for the 20th Attribute

Figure 22 shows the 5 box plots for the 20th attribute. We can see the oversampling generated slightly narrow boxes.

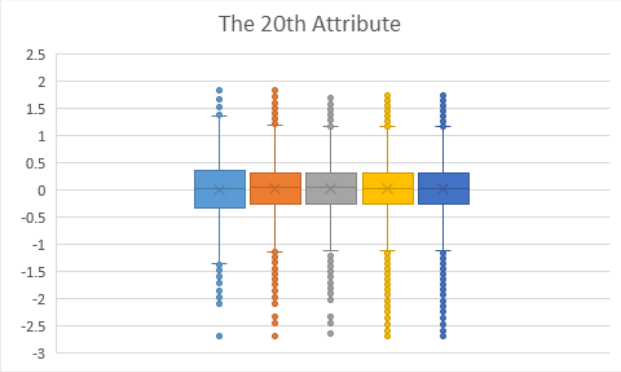

Fig. 22: Box plots for the 20th attribute in the original data, the data of 10-fold CV and oversampled only at the oversampling rate of 800%, and the data of 10-fold CV and oversampled only at the oversampling rate of 7200% from left to right respectively

A t-test for the mean was carried out on the data for the 20th attribute at the oversampling rate of 7200%. From the test equal variance was not assumed, because F=152.890 with significance=0.000 for the data. Table 31 shows the result of the t-test for the oversampled data at the oversampling rate of 7200%.

Table 31. The result of the t-test for the oversampled data at the oversampling rate of 7200% for the 20th attribute

Data at	Mean	SD	t	p
Original	0.0000	0.5465	-1.394	0.163
Over-sampled	0.0152	0.4567		

Because p=0.163 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 the 20th attribute are good at the oversampling rate of 7200%.

4.1.4 Random Forests after Dropping the 11th Attribute

In the t-test in the previous section 4.1.2, the oversampled data of attribute 11 was judged to belong to a different group from the original data value of the attribute, so attribute 11 was dropped and random forests were generated. Table 32 shows the random forests with an equal oversampling rate of 7200% in 10-fold cross-validation after dropping attribute 11.

Table 32. The result of random forests of the ozone data set with the equal oversampling rate of 7200% for each class after dropping attribute 11

Accuracy in 10-fold CV (%)	Confusion matrix		No. of Misclassified
99.8103	173298	4	351
	347	11333	

If we compute the TP rate for Table 30, the TP rate for ozone day is 0.97, while the TP rate for no ozone day is 1.0. Comparing Table 30 with Table 10 which is the result without dropping attribute 11, we have 2 more misclassifications for ozone day. When we consider ozone day belongs to positive, the precision = 11333/(11333+347) = 0.97 and recall = 11333/(11333+4) = 1.0 in Table 30. Note that the original paper that analyzed the ozone data has the precision and recall ranges of 0.4 ~ 0.6 only.

Table 33 shows the random forests when oversampled data only is used for training and the original data is used for testing after dropping the attribute 11.

Table 33. The result of random forests of the ozone data set with the equal oversampling rate of 7200% for each class when oversampled data only is used for training after dropping attribute 11.

Accuracy when the original data is used for testing(%)	Confusion matrix		No. of Misclassified
100	2374	0	0
	0	160	

Note that the TP rate for both classes is 1.0, so dropping misleading oversampled data such as attribute 11 generated a better result. The precision and recall values are 1.0.

5 Conclusion

A certain concentration of ozone on the ground can cause health problems, so accurate ozone level warnings are very important to prevent health problems. On the other hand, the characteristics of the public data related to ozone level detection in the UCI machine learning repository have some difficulty in the creation of an accurate machine learning model. The dataset has 72 relatively large numerical attributes and has been measured and collected for 7 years, and the distribution of ozone days and normal days is very unbalanced—ozone days are only 6.3% level in the dataset among around 2500 data entries, and the second difficulty is that the dataset might have multicollinearity

between the 72 conditional features in the original dataset, and there are some missing data also which is common in real-world data. Therefore, it was very difficult to create an accurate classification model for ozone days until now.

In this paper to solve the high dimensionality and multicollinearity of the dataset, PCA is applied first. As a result, the 72 attributes are reduced to 20 attributes and generated a slightly better machine learning model of random forests. So, we needed more sufficient data, especially for the minor class. For this purpose oversampling method of SMOTE which is one of the representative oversampling methods was applied at very high rates repeatedly to find enough size of samples for better classification models.

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 after applying oversampling at the same very high rate for each class and generating much more synthetic data than the original data and using it for training the random forests. However, because such synthetic data by oversampling is different from the original data, the synthetic data is compared with the original data to see if it is statistically reliable using boxplot and t-test. As shown in Table 32 and Table 33, the results of the experiment showed that when the oversampling rate was relatively high with the suggested method as we can see in the experiment, it could be possible to generate synthetic data with statistical characteristics similar to the original data on the condition that statistical tests are backed, 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, from 94% to 100%. Note that the random forests generated from the original data alone have no capability of classifying ozone days as we see in Table 2.

Conventionally until now, oversampled data has been used to train machine learning models neglecting statistical analysis, so, we may wonder how much it resembles the original data so that the synthetic data may be used to improve machine learning models. Applying a similar approach to this study to wine data in the UCI machine learning repository also showed that this approach was very useful in generating a very accurate knowledge model of random forests, [20]. Therefore, this paper is significant in the sense that it shows how we can apply a statistical methodology to test how reliable the oversampled data can be and it also shows this method can be effective in dealing with class imbalance problems. Note that the ozone day data were collected every day for 7 years, so future research will be building a knowledge model that can predict whether or not ozone day is in a few days based on the time-series data.

Acknowledgment:
This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2024.

References:

[1] K. Zhang, W. Fan, X. Yuan, I. Davidson, Forecasting Skewed Biased Stochastic Ozone Days: Analyses and Solutions, *Knowledge and Information Systems*, Vol. 14, 2008, pp. 299-326. https://doi.org/10.1007/s10115-007-0095-1.

[2] W. Jia, M. Sun, J. Lian, Feature dimensionality reduction: a review, *Complex Intelligent Systems*, Vol. 8, 2022, pp. 2663-2693. https://doi.org/10.1007/s40747-021-00637-x.

[3] S. Saha, S. Bhattacharya, A Survey: Principle Component Analysis(PCA), *International Journal of Advanced Research in Science and Engineering*, Vol. 6, Issue 6, 2017, pp. 312-320.

[4] Y. Wei, Y. Tang, P.D. McNicholas, Flexible High-Dimensional Unsupervised Learning with Missing Data, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 3, 2020, pp. 610-621.

[5] A. Sarkar, S.S. Ray, A. Prasad, C. Pradhan, A Novel Detection Approach of Ground Level Ozone using Machine Learning Classifiers, *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud),* Palladam, India, 11-13 November 2021, DOI: 10.1109/I-SMAC52330.2021.9640852.

[6] V. Laveglia, E. Trentin, Downward-Growing Neural Networks, *Entropy*, Vol. 25, No. 5, 733, 2023. https://doi.org/10.3390/e25050733.

[7] J. Shlens, A Tutorial on Principle Component Analysis, arXiv:1404.1100**,** https://doi.org/10.48550/arXiv.1404.1100. (Accessed Date: June 24, 2024).

[8] H. Almuallim, S. Kaneda, Y. Akiba, Development and Applications of Decision Trees, *Expert Systems*, edited by C.T.

Leondes, Academic Press, Vol. 1, 2002, pp. 53-77.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] K. Zhang, W. Fan, X. Yuan, Ozone Level Detection, *UCI Machine Learning Repository,* 2008, https://doi.org/10.24432/C5NG6W. (Accessed Date: June 1, 2024).

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

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

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

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

[20] H. Sug, An Oversampling Technique with Descriptive Statistics, *WSEAS Transactions on Information Science and Applications*, Vol. 21, 2024, pp. 318-332.

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

The sole author contributed to 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 2024.

Conflict of Interest

The author has no conflicts of interest to declare that are relevant to the content of this article.

APPENDIX

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

Attribute name	Value Range	D. values	Mean	SD
WSR0	0 ~ 7.5	68	1.64	1.272
WSR1	0 ~ 7.7	70	1.586	1.267
WSR2	0 ~ 7.1	65	1.546	1.24
WSR3	0 ~ 7.3	66	1.526	1.206
WSR4	0 ~ 7.2	64	1.523	1.199
WSR5	0 ~ 7.4	63	1.542	1.172
WSR6	0 ~ 7.4	66	1.638	1.162
WSR7	0 ~ 7.5	67	2.047	1.161
WSR8	0.1 ~ 9.2	69	2.539	1.185
WSR9	0.1 ~ 8.5	70	2.848	1.221
WSR10	0 ~ 8.7	76	2.97	1.302
WSR11	0.1 ~ 8.8	77	3.016	1.386
WSR12	0 ~ 9	77	3.044	1.418
WSR13	0 ~ 9.6	78	3.107	1.44
WSR14	0.1 ~ 9.1	77	3.178	1.423
WSR15	0 ~ 8.9	78	3.231	1.372
WSR16	0.2 ~ 8.7	72	3.193	1.282
WSR17	0.2 ~ 8.1	73	2.935	1.232
WSR18	0 ~ 7.5	70	2.561	1.239
WSR19	0 ~ 7.1	65	2.286	1.214
WSR20	0 ~ 8.7	68	2.09	1.205
WSR21	0 ~ 9.3	69	1.938	1.208
WSR22	0 ~ 7.7	68	1.804	1.233
WSR23	0 ~ 8.3	65	1.709	1.263
WSR_PK	0.8 ~ 9.6	74	4.172	1.174
WSR_AV	0.4 ~ 6.4	55	2.315	0.923
T0	-1.8 ~ 29.9	282	18.649	7.021
T1	-2.1 ~ 29	284	18.348	7.087
T2	-2.6 ~ 28.8	287	18.061	7.154
T3	-2.8 ~ 28.3	283	17.821	7.205
T4	-3.2 ~ 28.1	283	17.611	7.254
T5	-3.6 ~ 28.2	291	17.476	7.312
T6	-3.2 ~ 28.7	295	17.589	7.518
T7	-2.8 ~ 30.1	311	18.418	7.872
T8	-1.9 ~ 31.4	313	19.779	7.879
T9	-1.2 ~ 33.8	314	21.217	7.759
T10	-1.2 ~ 36.4	327	22.463	7.693
T11	-0.3 ~ 38.5	330	23.394	7.628
T12	0.3 ~ 40.4	334	24.025	7.561
T13	0.9 ~ 41.3	335	24.433	7.466
T14	1.5 ~ 41.6	335	24.705	7.385
T15	1.7 ~ 41.3	339	24.72	7.303
T16	0.6 ~ 41.1	337	24.398	7.237
T17	-0.6 ~ 39.9	329	23.632	7.179
T18	-0.2 ~ 37.8	321	22.51	7.087
T19	0.1 ~ 36.1	306	21.426	6.923
T20	0.2 ~ 34.6	302	20.615	6.866
T21	-0.3 ~ 33.4	294	20.032	6.864
T22	-1.4 ~ 32.6	287	19.503	6.901
T23	-1.2 ~ 31.3	284	19.062	6.961
T_PK	1.7 ~ 41.6	330	25.578	7.152
T_AV	0.3 ~ 33.6	296	20.84	7.013
T85	-7.1 ~ 24.5	251	13.575	4.874

Attribute name	Value Range	D. values	Mean	SD
RH85	0.01 ~ 1	100	0.577	0.258
U85	-15.77 ~ 18.56	1288	2.136	4.725
V85	-18.1 ~ 22.16	1461	1.662	6.134
HT85	1351 ~ 1642	368	1531.494	36.695
T70	-9.9 ~ 16.2	245	5.931	3.867
RH70	0.01 ~ 1	100	0.406	0.268
U70	-14.37 ~ 28.21	1537	5.46	6.676
V70	-23.68 ~ 25.54	1429	0.994	6.186
HT70	2919 ~ 3249	441	3145.421	49.175
T50	-24.8 ~ -1.7	186	-10.511	3.882
RH50	0.01 ~ 1	100	0.305	0.249
U50	-14.92 ~ 42.36	1687	9.872	9.531
V50	-25.99 ~ 30.42	1509	0.83	7.355
HT50	5480 ~ 5965	85	5818.821	78.18
KI	-56.7 ~ 42.05	1047	10.511	20.718
TT	-10.1 ~ 59.15	657	37.388	11.23
SLP	9975 ~ 10350	71	10164.198	52.421
SLP_	-135 ~ 140	56	-0.12	35.829
Precp	0 ~ 20.65	174	0.372	1.318
Class	2 class values (0, 1)			