Cumhuriyet Dental Journal, 26(4): 374-380, 2023
DOI: https://doi.org/10.7126/cumudj.1332452

Cumhuriyet Dental Journal

| cdj.cumhuriyet.edu.tr | Founded: 1998 Available online, ISSN: 1302-5805
e-ISSN: 2146-2852 Publisher: Sivas Cumhuriyet Üniversitesi

An Oversampling Technique for Handling Imbalanced Data in Patients with Metabolic Syndrome and Periodontitis

S. Merve Altıngöz[1-a*], Batuhan Bakırarar[2-b], Elif Ünsal[3-c], Şivge Kurgan[3-d], Meral Günhan[3-e]

[1] *Department of Periodontology, Faculty of Dentistry, Lokman Hekim University, Ankara, Turkiye.*
[2] *Department of Biostatistics, Faculty of Medicine, Ankara University, Ankara, Turkiye.*
[3] *Department of Periodontology, Faculty of Dentistry, Ankara University, Ankara, Turkiye.*

**Corresponding author*

Research Article

History

Received: 25/07/2023
Accepted: 29/11/2023

ABSTRACT

Objectives: Periodontitis has been suggested to be associated with several systemic diseases and conditions including obesity, metabolic syndrome, diabetes, chronic renal disease, respiratory disorders, and cardiovascular diseases. Metabolic syndrome (MetS) is a collection of impairment and is a risk factor for type 2 diabetes and cardiovascular disease. Our study is aimed to handle MetS unbalanced data using the synthetic minority over-sampling technique (SMOTE) to increase accuracy and reliability.

Materials and Methods: Six metabolic syndrome patients and 26 systemically healthy subjects with periodontitis were recruited in this study. Clinical parameters (Plaque index (PI), gingival index (GI), probing pocket depth (PPD), clinical attachment loss (CAL), and bleeding on probing (BOP)) were obtained, smoking status and body-mass index (BMI), systemic diseases, fasting glucose levels, hemoglobin A1c (HbA1c) levels and serum advanced glycation end-products (AGE) levels were recorded by one examiner. First, the data was pre-processed by removing missing values, outliers and normalizing the data. Then, SMOTE technique was used to oversample the minority class. SMOTE works by creating synthetic data points that are similar to the existing minority class instances. The experimental dataset included numerous machine learning algorithms and assessed accuracy using both pre- and post-oversampling methods.

Results: Our findings suggest that by increasing the sample size of a study, researchers can gain more accurate and reliable results. This is especially important when studying a population with a lower sample size, as the results may be skewed.

Conclusions: SMOTE may result in over fitting on numerous copies of minority class samples.

Key Words: Chronic Periodontitis; Metabolic Syndrome; Over-Sampling; Synthetic Minority Over-Sampling Technique.

[a] *smerve.unal@yahoo.com* *https://orcid.org/0000-0002-9709-6226*
[b] *batuhan_bakirarar@hotmail.com* *https://orcid.org/0000-0002-5662-8193*
[c] *unsal.e@gmail.com* *https://orcid.org/0000-0002-7843-6088*
[d] *sivgeakgun@gmail.com* *https://orcid.org/0000-0002-7868-4217*
[e] *meralgunhan@yahoo.com* *https://orcid.org/0000-0002-3848-6195*

How to Cite: *Altıngöz SM, Bakırarar B, Ünsal E, Kurgan Ş, Günhan M. (2023) An Oversampling Technique for Handling Imbalanced Data in Patients with Metabolic Syndrome and Periodontitis, Cumhuriyet Dental Journal, 26(4):374-380.*

Introduction

Periodontitis (PD) is a chronic inflammatory condition that affects supportive tissues and is linked to periodontal pathogens.[1] Previous studies have shown positive association between periodontitis and systemic alterations such as low-grade glucose intolerance, inflammation[2], insulin resistance[3], a pro-coagulant state[4], endothelial dysfunction[5] and vascular dysfunction.[6] Some other studies have indicated a possible relationship between lipid abnormalities, arterial blood pressure and periodontitis.[7, 8]

Metabolic syndrome (MetS) is a condition with many risk factors including impaired glucose tolerance, hypertension, abdominal obesity, atherogenic dyslipidemia, insulin resistance and dyslipidemia that predispose to cardiovascular disease and diabetes mellitus.[9] National Cholesterol Education Program Adult Treatment Panel III (NCEP ATP III) diagnosis criteria for metabolic syndrome includes the presence of three or more of the following abnormal states: atherogenic dyslipidemia, hypertension, abdominal obesity, and hyperglycemia/insulin resistance.[10]

Persistent low-grade inflammatory state in periodontal disease may cause the progression of insulin resistance and the imbalance of the interaction between cytokine and periodontium.[11, 12] Chronic systemic inflammation in individuals with periodontal disease may predispose to the development of MetS components or vice versa.[13] The most common limitation of the studies investigating this relationship in literature is the insufficient number of cases. The test and control groups are unbalanced due to a lack of participants in the case group. In statistical analyses, this scenario leads to type 2 errors and statistical insignificance.

It is necessary to handle the imbalance of the dataset before training a model. There are numerous methods to handle the imbalance of the datasets (e.g., oversampling, under-sampling, or a combination of both). Oversampling can be categorized as Random Oversampling and

Synthetic Oversampling. Random oversampling replicates the minority class samples and balances the data without loss of information.[14] Due to the replication of the same data, the dataset is prone to overfitting. At this point, to balance the dataset Synthetic Minority Oversampling Technique (SMOTE) generates new synthetic samples using the k-nearest neighbor technique. SMOTE works by creating new synthetic examples that are similar to existing minority class examples, but the new synthetic examples may not reflect the accurate distribution of the data.[15] This means that if the same minority class examples are used repeatedly to generate new synthetic examples, then the model may become overly familiar with those examples, resulting in overfitting.

In this study, the oversampling method was used to avoid reduction to the group with lower sample size. With this method, the sample size of the group with fewer samples was increased by using all features of the data set, and thus the imbalance in the groups was eliminated. The descriptors of the initial version and the increased sample size data set which was simulated with oversampling method showed that there was no bias in the data set in terms of variables. As a final process, data mining methods were applied for the original data set and the data set simulated with oversampling method and it was shown that the results were significant for the simulated data.

Patients with small sample sizes, such as those with rare disorders or study conducted during unusual times like pandemics, may face difficulties in accessing patients and obtain their data.

To address this issue, researchers can use alternative methods such as virtual studies, online surveys, and data mining from existing databases. Additionally, researchers can collaborate with other research groups to pool data and increase sample size.

The goal of this study is to keep the properties of the dataset that are rare or have small sampling size for various reasons, and to acquire reliable results by increasing the number of cases enrolled, hence boosting the study's power, and statistically preventing type 2 mistake. To achieve this goal, the study will employ a variety of methods, such as oversampling, where oversampling involves randomly duplicating instances of the minority class in the dataset to balance out the class distribution. Synthetic data generation involves creating new data points based on existing data points in the dataset.

Materials and Methods

Patients

Six metabolic syndrome patients with periodontitis (MetS-P group) and twenty-six systemically healthy individuals with periodontitis (H-P group) were included in the study who applied to the Department of Periodontology, School of Dentistry, and the Department of Endocrinology, School of Medicine, at Ankara University. The periodontal diagnosis was based on the World Workshop on Periodontal and Peri-Implant Diseases and Conditions (2017).[16] The periodontitis group included patients with Stage III generalized periodontitis. The study was approved by the Ankara University Faculty of Dentistry Human Research Ethics Committee, (No: 9/1, on 08.01.2011) and was conducted in consistency with the Helsinki Declaration. All participants were asked to sign the informed consent form.

Inclusion criteria (for MS group):

1) fasting glucose level <126 mg/dL and >100 mg/dL

2) 2-hour glucose tolerance test glucose level <200 mg/dL;

3) HbA1c >5.7% and <6.4%.

Exclusion criteria for all groups:

Pregnant and lactating mothers, obese individuals, patients with systemic disorders including diabetes, cardiovascular disease, and cancer, or antibiotics or anti-inflammatory drug usage, or medication with calcium channel blockers, phenytoin, or cyclosporine, or patients who had received non-surgical periodontal therapy in the previous 6 months and surgical periodontal therapy in the previous 12 months all excluded.

Smokers of 10 or more cigarettes per day for at least the last five years were defined as smokers. Individuals who never smoked or quit smoking at least two years ago were defined as non-smokers.

Evaluation of Periodontal Status and Metabolic Syndrome

Clinical periodontal parameters including plaque index (PI)[17], gingival index (GI)[18], probing pocket depth (PPD), clinical attachment loss (CAL), and percentage of bleeding on probing sites (BOP%) were recorded from six different sites for all teeth except third molars at first visit in the Department of Periodontology and Department of Endocrinology.

Body mass index, fasting glucose (mg/dl; hexokinase method DXC, Beckman Coulter) and HbA1c (%; HPLC kit, Immuchrome GmbH, Germany) were recorded; waist and hip circumferences were measured, and waist-hip ratio was calculated for all participants.

Saliva and Serum samples

Unstimulated saliva samples were obtained after a 12-hour fasting period, in the early hours of the day. Whole saliva (2 ml) was collected using sterile polypropylene tubes. The samples were centrifuged at 10000 ×g for 10 minutes.[19]

Blood samples were collected from antecubital vein by venipuncture. The samples were centrifuged at 4000 ×g for 10 minutes. Saliva and blood samples were preserved at -80°C until the day of the analysis.[20]

Measurement of Advanced Glycation End-products (AGE)

AGE was determined using OxiSelect™ and Oxiselect™ AGE Adduct kits (Cell Biolabs, Inc. San Diego, CA, USA), respectively. ELISA tests were conducted according to the manufacturer's instructions. SynergyTM HT Microplate Reader (Bio-Tek Instruments, Winooski, WT, USA) was used to measure the color change at 450 nm.

Synthetic Minority Oversampling Method

It is not logical to decrease the sample size of the group which has higher number of patients when the minority class sample size is too small. In our study Synthetic Minority Oversampling Technique (SMOTE) method was used for small minority class samples. The SMOTE algorithm uses an oversampling strategy to rebalance the initial training set which introduces synthetic examples instead of just replicating instances of the minority class. In the medical disciplines, imbalanced data is a concern because there are permanent data for patients more than healthy data. The SMOTE is a method to create artificial instances by operating in "feature space" as opposed to "data space." This lessens the risk of overfitting and enhances the accuracy of the results.[21]

The steps of SMOTE algorithm are:

1. For each minority instance, k number of nearest neighbors are found;

$$k = (SMOTE\%)/100$$

2. One of these neighbors are chosen and a synthetic point is placed anywhere on the line joining the point under consideration and its chosen neighbor

3. A new synthetic sample is generated by interpolating between the selected minority class sample and the randomly selected neighbor.[15]

With this method, the sample size of the minority class was increased by using all the features of the data set, and thus the imbalanced problem in the groups was eliminated. With the initial version of the data set and the oversampling method, the descriptors of the increased sample size were given, and it was shown that the data set did not cause any bias in terms of variables. Finally, data mining methods have been applied for the original data set and the oversampling replicated data set, and the results have been shown to be better for the replicated data.

Statistical Analyses

The data were evaluated using WEKA 3.7 and SPSS 11.5 software. The number of patients (percent) for qualitative variables; mean ± standard deviation and median (minimum-maximum) for quantitative data were utilized as descriptors. The Mann-Whitney U test was used because there was no statistically significant difference between the categories of the two-category qualitative variable in terms of the quantitative variable and there was no assumption about the normal distribution. Chi-square and Fisher-exact tests were used to determine the association between the two qualitative variables. All tests have been conducted with a significance level of α=0.05. In WEKA program, Multilayer Perceptron, J48 and Support Vector Machine, which are Classification methods, were used. While 10-fold Cross Validation test option was used for evaluating the data set; Accuracy, F-Measure, MCC and Precision-Recall Curve (PRC Area) were used as data mining performance criteria.

Results

Since there are so many variables in the data set, the importance of the variables and the values added to the data set were examined using the Info Gain Attribute Eval, Gain Ratio Attribute Eval and Chi-Squared Attributed Eval methods in WEKA and variables, that were determined to be insignificant by the three methods and considered to be of essential importance for clinical information were excluded from the dataset. The data set contained a total of 10 variables: 9 independent variables and 1 dependent variable. These variables are gender, smoking, BMI, HbA1c, hypertension, other disease, other medication, PD, serum levels of AGE and metabolic syndrome. Percentages of variable importance of 9 independent variables according to metabolic syndrome, which is the dependent variable, are given in Figure 1.

Twenty-six periodontally healthy MetS patients (H-P group; 10 female and 16 males), 6 MetS patients with periodontitis (MetS-P group; 4 female and 2 males) were the original dataset. Analysis results of the original data set are presented in Table 1. Three patients among MetS-P group, 15 patients among H-P group were smokers. BMI, HbA1c level and having other medications were significantly higher in MetS-P group ($p<0.05$). No significant differences were found in terms of hypertension and other diseases between MetS-P and H-P group ($p>0.05$). While PD was significantly higher in H-P group; serum levels of AGE were significantly higher in H-P group. The difference was statistically significant between group regarding BMI, HbA1c, other medications, PD, and serum levels of AGE (respectively p=0.018, p=0.012, p=0.038, p=0.010, p=0.031).

Analysis results of the original data set is presented in Table 2. Thirteen patients among MetS-P group were smokers. BMI, HbA1c level and having other diseases and other medications were significantly higher in MetS-P group compared to systemically healthy group ($p<0.001$). No significant difference was found between MetS-P and H-P group in terms of hypertension ($p>0.05$). While PD was significantly higher in H-P group; serum levels of AGE were significantly higher in H-P group. The difference was statistically significant between MetS-P and H-P groups in terms of serum variables of gender, BMI, HbA1c, other diseases, other medications, PD, serum AGE levels (respectively p=0.026, p<0.001, p<0.001, p<0.001, p<0.001, p<0.001, p<0.001).

When Tables 1 and 2 were compared, it was seen that gender and other diseases, which had not been significantly different in Table 1, became significant in Table 2. This suggests that the presence of other diseases may influence the gender differences in the prevalence of the disease. In addition, it was seen that the p values of other variables that were significant in Table 1 also increased in significance. In Table 2, it was observed that all identifiers/properties of the variables were preserved from Table 1 and that the p-values for each variable were slightly lower than in Table 1. This indicates that the variables in Table 2 are slightly more reliable than those in Table 1. This suggests that the variables in Table 2 are more reliable and should be considered when making decisions.

In Table 3, performance criteria of data mining methods for real and simulated data sets were given. In the real data set, it is seen that the performance criteria of the MetS group were lower than systemically healthy group. The reason for this was that the sample size of this group was less than the other group. In the simulated data set, when the sample number of this group was made equal to the H-P group, increases in the performance criteria were observed. For example, while the accuracy value for the multilayer perceptron method in the real data set was 0.667 in the MetS-P group, this value increased to 0.923 in the simulated data set. The reason for this was that data mining methods, like basic statistical methods, were also affected by the uneven distribution between groups and the number of samples. When this value was close to 1 for all performance measures, that made the classification more successful.

Discussion

The objective of the study was to use the oversampling method to increase the amount of data in the MetS-P group (n=6) and balance it with the data set in the H-P group (n=26). In addition to the gender and other disease factors, which were not *p*-significant, balancing the data sets revealed an increase in the existing *p*-significance values. The oversampling method was utilized for the first time in dentistry with study.

El-Sayed *et al.* (2015), in their study on 100 autistic and 15 non-Autism individuals, performed SMOTE to 15 non-Autistic individuals, increased the number of this group to 60 and avoid the imbalance between groups. They applied Support Vector Machine, J48, Naive Bayes and Multilayer Perceptron, which are data mining algorithms, were performed to the balanced groups and the results were compared. The results showed that there was an improvement in performance criteria.[21]

Shin *et al.* (2020), used the SMOTE to eliminate the strong imbalance between the healthy group and depression group. Cross-validation approach was used while creating the model, and the current data set was randomly divided into three different data sets, and each was considered independently.[22]

Ramezankhani *et al.* (2016) balanced the minority class with oversampling method, due to imbalance of the diabetic and non-patient group totally including 6647 individuals (1st group 729 patients, 2nd group 5918 patients). They created 100%, 200%, 300%, 400%, 500%, 600% and 700% training sets from the original data and as a result, they showed that the data group with 700% oversampling rate had the best performance.[23]

Fotouhi *et al.* (2019) used data analysis in cancer diagnosis in their study. However, unbalanced data distribution between classes caused erroneous interpretation of the results. Incorrect results to be obtained because of false evaluation may cost patients' lives. Therefore, it is very important in medicine to solve the class imbalancement problem. Fotouhi conducted a study on the results of the unbalanced data problem and compared the results of these methods by considering the undersampling and oversampling methods. They used RIPPER, MLP, KNN and C4.5 as data mining method and reported the best method for each combination obtained. They used AUC as a performance criterion and emphasized that the performance of their classifiers for different unbalanced cancer datasets improved in 90% of cases when the data became balanced. They also reported that oversampling methods have better results than undersampling methods.[24]

Nguyen *et al.* (2019) studied on the data set which were obtained from 9948 patients' records for, 1904 were diagnosed with Type 2 Diabetes. The diabetes estimation for 2012 is based on data from prior years. (2009–2011). Using the SMOTE method for the unbalanced class in the data set, the data was produced, and the data set became 1: 2 and 1: 1. They compared performance criteria by applying the data mining algorithm they developed for the new data sets and the original data set. They emphasized that the use of 150% and 300% SMOTE did not improve AUC, but resulted in an increase in sensitivity (49.40% and 71.57%, respectively).[25]

In the study performed by Cui *et al.* (2019), a total of 230 variables were examined in samples obtained from 106 lung cancer patients. In the study, SMOTE was applied to cope with unbalanced and small data sets in radiotherapy toxicity modeling, they concluded that there was an improvement in the results and that it was more appropriate to interpret the data performed with SMOTE.[26]

In rare diseases or extraordinary situations such as COVID-19 pandemic, the number of the collected data may not be balanced. The purpose of the oversampling method is to make the group with a small number of data sets closer to the group with a larger number of data sets, that is, to make the data sets balanced. Thus, it will be possible to interpret the data. There are many studies in the literature in which the oversampling method is applied in the field of medicine/health. While searching the literature, it was seen that in some data sets one group had quite high n values (Ramazankhi *et al.* (2016), 1st group 729 patients, 2nd group 5918 patients). In some studies, on the other hand, it was observed that there were smaller sample sizes in all groups in data sets (El-Sayed *et al.* (2015), 100 autistic and 15 non-Autism individuals). The reason for this difference; the high data groups created by scanning the previously recorded data and the lower data groups created by recording one-to-one in the clinic. Since our study consisted of data recorded in the clinical setting, the n number was obtained relatively lower.

When applying the oversampling method, there is no numerical lower limit in the data sets. However, researchers should keep the data numbers as high as possible when planning studies. This method is a reliable method that can be used not for studies where standard data can be obtained, but for duplication of data obtained in small numbers in pandemics and rare diseases. The purpose of the method is not to obtain false data, but to increase the value of existing scientific data.

Similarly, there were changes in data mining performance criteria when comparing the real and simulated data sets in Table 3. In terms of the study's significance, using simulation approaches to generate data for rare and significant groups is critical, as shown in our study.

As can be shown in our study, completing the data in the lower sample size groups eliminates the statistical significance resulting from the sample size for the clinically known factors and makes the study more beneficial by retaining the distribution and properties of the data. Consequently, the study provides both clinical and statistical significance.

Conclusions

Researchers should consider the most appropriate methods for collecting data, such as surveys, interviews, or focus groups. They should also consider the most appropriate way to analyze the data, such as statistical analysis or qualitative analysis. The purpose of researchers should be to use the data obtained to answer questions, draw conclusions, and make recommendations. This could include developing new treatments, understanding the causes of a disease, or identifying risk factors for a particular condition. Researchers should also strive to ensure that their research is conducted ethically. In extreme cases, such as pandemics, it may be impossible to include the number of patients who should be included in the research when designing scientific studies. The reality of Covid has reminded us of what we need to reconsider while doing our work in the scientific field, as well as in all aspects of our lives.

The goal of this study's planning is not to demonstrate that this can be used as an alternative method in every situation. This is how we can use statistical science in scientific studies, where it is a valuable alternative method that may bring a different approach that can be used not to miss cases for special conditions and diseases. Researchers should keep in mind that power analysis will normally be conducted before the data collection while designing a study and calculating statistical power is a crucial step in determining the sample size. Our study intends to demonstrate how this method works in a study group that we expect will be easily followed by everyone, rather than establishing the association between metabolic syndrome and periodontitis.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. Pihlstrom BL, Michalowicz BS, Johnson NW. Periodontal diseases. Lancet 2005, 366, 1809–1820.

2. Slade GD, Offenbacher S, Beck JD, et al. Acute-phase inflammatory response to periodontal disease in the US population. J Dent Res. 2000;79:49-57.

3. Saito T, Shimazaki Y, Kiyohara Y, et al. The severity of periodontal disease is associated with the development of glucose intolerance in non-diabetics: The Hisayama study. J Dent Res. 2004;83:485-490.

4. Taylor BA, Tofler GH, Carey HM, et al. Full-mouth tooth extraction lowers systemic inflammatory and thrombotic markers of cardiovascular risk. J Dent Res. 2006;85:74-78.

5. Higashi Y, Goto C, Jitsuiki D, et al. Periodontal infection is associated with endothelial dysfunction in healthy subjects and hypertensive patients. Hypertension. 2008;51:446-453.

6. Tonetti MS, D'Aiuto F, Nibali L, et al. Treatment of periodontitis and endothelial function. N Engl J Med. 2007;356:911-920.

7. Katz J, Flugelman MY, Goldberg A, et al. Association between periodontal pockets and elevated cholesterol and low density lipoprotein cholesterol levels. J Periodontol. 2002;73:494 500.

8. Losche W, Karapetow F, Pohl A, et al. Plasma lipid and blood glucose levels in patients with destructive periodontal disease. J Clin Periodontol. 2000;27:537-541.

9. International Diabetes Federation. The IDF Consensus Definition of the Metabolic Syndrome in Children and Adolescents, 2007.

10. Ford ES, Giles WH, Mokdad AH. Increasing prevalence of the metabolic syndrome among U.S. adults. Diabetes Care 2004, 27, 2444–2449.

11. Makkar H, Reynolds MA, Wadhawan A, Dagdag A, Merchant AT, Postolache TT. Periodontal, metabolic, and cardiovascular disease: exploring the role of inflammation and mental health. Pteridines.2018;29:124-163.

12. Grundy SM, Cleeman JI, Daniels SR, et al. Diagnosis and management of the metabolic syndrome: An American Heart Association/National Heart, Lung, and Blood Institute scientific statement. Cardiol Rev. 2005;13:322-327.

13. Nibali L, D'Aiuto F, Griffiths G, Patel K, Suvan J, Tonetti MS. Severe periodontitis is associated with systemic inflammation and a dysmetabolic status: a case-control study. J Clin Periodontol.2007;34:931-937.

14. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Min. Knowl. Disc. 28(1), 92–122 (2014).

15. Fernndez A, Garca S, del Jesus MJ, Herrera F. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets, Fuzzy Sets and Systems, 159(18), 23782398, 2008.

16. Tonetti MS, Greenwell H, Kornman KS. Staging and grading of periodontitis: framework and proposal of a new classification and case definition. J Periodontol. 2018; 89(Suppl 1): 159-172.

17. Silness J, Loe H. Periodontal disease in pregnancy. II. Correlation between oral hygiene and periodontal condition. *Acta Odontol Scand.* 1964;22:121–135.

18. Loe H, Silness J. Periodontal disease in pregnancy. I. Prevalence and Severity. *Acta Odontol Scand.* 1963;21:533–551.

19. Caglayan F, Miloglu O, Altun O, et al. Oxidative stress and myeloperoxidase levels in saliva of patients with reccurrent aphthous stomatitis. *Oral Dis.* 2008;12:700–704.

20. Tayman MA, Kurgan Ş, Önder C, Güney Z, Serdar MA, Kantarcı A, Günhan M (2019) Affiliations expandA disintegrin-like and metalloproteinase with thrombospondin-1 (ADAMTS-1) levels in gingival crevicular fluid correlate with vascular endothelial growth factor-A, hypoxia-inducible factor-1α, and clinical parameters in patients with advanced periodontitis. J Periodontol 90(10):1182–1189.

21. El-Sayed AA, Mahmood MAM, Meguid NA, Hefny HA. Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE), *2015 Third World Conference on Complex Systems (WCCS)*, Marrakech, 2015, pp. 1-5.

22. Shin D, Lee KJ, Adeluwa T, Hur J. Machine Learning-Based Predictive Modeling of Postpartum Depression. *Journal of clinical medicine*, 2020. *9*(9), 2899.

23. Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F, Khalili D. The Impact of Oversampling with SMOTE on the Performance of 3 Classifiers in Prediction of Type 2 Diabetes. *Medical decision making:an international journal of the Society for Medical Decision Making*, 2016, *36*(1), 137–144.

24. Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics*, 2019, *90*, 103089.

25. Nguyen BP, Pham HN, Tran H, Nghiem N, Nguyen QH, Do T T T, et al. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Computer methods and programs in biomedicine*, 2019, *182*, 105055.

26. Cui S, Luo Y, Tseng HH, Ten Haken RK, El Naqa I. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. *Medical physics*, 2019, *46*(5), 2497–2511.

Table 1. Original dataset statistics

Variables		Groups		
		MetS-Periodontitis	Healthy-Periodontitis	p value
Gender, n (%)	Female	4 (66.7)	10 (38.5)	0.365a
	Male	2 (33.3)	16 (61.5)	
Smoking, n (%)	Smoker	3 (50.0)	15 (57.7)	1.000a
	Non-smoker	3 (50.0)	11 (42.3)	
BMI, n (%)	Normal	2 (33.3)	8 (30.8)	0.018a
	Pre-obese	0 (0.0)	13 (50.0)	
	Obese-1	4 (66.7)	5 (19.2)	
HbA1c	Mean ± SD	5.62±0.88	4.81±0.32	0.012b
	Median (Min.-Max.)	5.80 (4.00-6.40)	4.80 (4.30-5.50)	
Hypertension, n (%)	No	5 (83.3)	26 (100.0)	0.188a
	Yes	1 (16.7)	0 (0.0)	
Other diseases, n (%)	No	3 (50.0)	21 (80.8)	0.148a
	Yes	3 (50.0)	5 (19.2)	
Other medications, n (%)	No	2 (33.3)	21 (80.8)	0.038a
	Yes	4 (66.7)	5 (19.2)	
PD	Mean ± SD	3.04±0.68	3.93±0.61	0.010b
	Median (Min.-Max.)	2.86 (2.34-4.07)	3.93 (2.85-4.97)	
AGE Serum	Mean ± SD	0.68±0.28	0.39±0.07	0.031b
	Median (Min.-Max.)	0.87 (0.35-0.90)	0.39 (0.28-0.61)	

Abbreviations: BMI, body-mass index; HbA1c, hemoglobin A1c; PD, pocket depth; AGE, advanced glycation end-products; SD, Standard deviation; Min, Minimum; Max, Maximum. a: Fisher-exact test. b: Mann-Whitney U test

Table 2. Statistics of the duplicated data set using the SMOTE method

Variables		Groups		
		MetS-Periodontitis	Healthy-Periodontitis	p value
Gender, n (%)	Female	18 (69.2)	10 (38.5)	0.026a
	Male	8 (30.8)	16 (61.5)	
Smoking, n (%)	Smoker	13 (50.0)	15 (57.7)	0.578a
	Non-smoker	13 (50.0)	11 (42.3)	
BMI, n (%)	Normal	9 (34.6)	8 (30.8)	<0.001a
	Pre-obese	0 (0.0)	13 (50.0)	
	Obese-1	17 (65.4)	5 (19.2)	
HbA1c	Mean ± SD	5.61±0.78	4.81±0.32	<0.001c
	Median (Min.-Max.)	5.80 (4.00-6.40)	4.80 (4.30-5.50)	
Hypertension, n (%)	No	22 (84.6)	26 (100.0)	0.110b
	Yes	4 (15.4)	0 (0.0)	
Other diseases, n (%)	No	12 (46.2)	21 (80.8)	0.010a
	Yes	14 (53.8)	5 (19.2)	
Other medications, n (%)	No	8 (30.8)	21 (80.8)	<0.001a
	Yes	18 (69.2)	5 (19.2)	
PD	Mean ± SD	3.04±0.61	3.93±0.61	<0.001c
	Median (Min.-Max.)	2.86 (2.34-4.07)	3.93 (2.85-4.97)	
AGE Serum	Mean ± SD	0.68±0.25	0.39±0.07	<0.001c
	Median (Min.-Max.)	0.87 (0.35-0.91)	0.39 (0.28-0.61)	

Abbreviations: BMI, body-mass index; HbA1c, hemoglobin A1c; PD, pocket depth; AGE, advanced glycation end-products; SD, Standard deviation; Min, Minimum; Max, Maximum. a: Fisher-exact test. b: Mann-Whitney U test

Table 3. Performance matrix table of data mining methods that includes original and replicated data set

Data set	Methods		Accuracy	F-measurement	MCC	PRC Area
Original data set	Multilayer Perceptron	MetS-Periodontitis	0.667	0.727	0.675	0.787
		Healthy-Periodontitis	0.962	0.943	0.675	0.986
		Overall	0.906	0.903	0.675	0.948
	J48	MetS-Periodontitis	0.667	0.615	0.520	0.589
		Healthy-Periodontitis	0.885	0.902	0.520	0.961
		Overall	0.844	0.848	0.520	0.891
	Support Vector Machine	MetS-Periodontitis	0.333	0.444	0.395	0.347
		Healthy-Periodontitis	0.962	0.909	0.395	0.860
		Overall	0.844	0.822	0.395	0.764
Simulated data set	Multilayer Perceptron	MetS-Periodontitis	0.923	0.906	0.808	0.973
		Healthy-Periodontitis	0.885	0.902	0.808	0.978
		Overall	0.904	0.904	0.808	0.975
	J48	MetS-Periodontitis	0.846	0.880	0.772	0.889
		Sağlıklı Periodontitis	0.923	0.889	0.772	0.850
		Overall	0.885	0.884	0.772	0.869
	Support Vector Machine	MetS-Periodontitis	1.000	0.945	0.891	0.897
		Healthy-Periodontitis	0.885	0.939	0.891	0.942
		Overall	0.942	0.942	0.891	0.919