International Journal of Applied Science and Engineering

Special issue: The 10th International Conference on Awareness Science and Technology (iCAST 2019)

Synthetic oversampling based decision support framework to solve class imbalance problem in smoking cessation program

Khishigsuren Davagdorj[1], Jong Seol Lee[1], Kwang Ho Park[1], Pham Van Huy[2], Keun Ho Ryu[2, 3*]

[1] Database and Bioinformatics Laboratory, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju, South Korea
[2] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[3] Department of Computer Science, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju, South Korea

ABSTRACT

Smoking is one of the significant avoidable risk factors for premature death. Most smokers make multiple quit attempts during their lifetime but smoking dependence is not easy and many people eventually failed quit attempts. Predicting the likelihood of success in smoking cessation program is necessary for public health. In recent years, a few numbers of decision support systems have been developed for dealing with smoking cessation based on machine learning techniques. However, the class imbalance problem is increasingly recognized as serious in real-world applications. Therefore, this paper presents a synthetic minority over-sampling technique (SMOTE) based decision support framework in order to predict the success of smoking cessation program using Korea National Health and Nutrition Examination Survey (KNHANES) dataset. We carried out experiments as follows: I) the unnecessary instances and variables have been eliminated, II) then we employed three variations of SMOTE, III) also the prediction models have been constructed. Finally, compare the prediction models to obtain the best model. Our experimental results showed that SMOTE improved the prediction performance of machine learning classifiers among evaluation metrics. Moreover, SMOTE regular based Random Forest (RF) and Naïve Bayes (NB) classifiers were determined the best prediction models in real-world smoking cessation dataset. Consequently, our decision support framework can interpret the important risk factors of smoking cessation using multivariate regression analysis.

Keywords: Smoking cessation; Risk factor analysis; Class imbalance; Synthetic minority oversampling; Machine learning classifiers.

1. INTRODUCTION

Over the past decades, the smoking epidemic is one of the biggest health threats the world has ever faced, filling more than 8 million people around the world each year. Although 1.2 million of those deaths are the result of non-smokers, it seems non-smokers being exposed to second-hand smokers (WHO, 2008). Smoking cessation can lead to a reduction in the risk of cancers such as lung, larynx, esophagus, and pancreas among smokers (Song et al., 2008). Essentially smoking is now well established as a perceived major cause of disease and early death, a dramatic rise of about 100 million deaths from the previous century and 1 billion estimated deaths during the 21st century. By 2030, the death toll is projected to reach more than 8 million per year. Furthermore, over 80% of smokers live in low and middle-income countries.

OPEN ACCESS

Received: March 29, 2020
Revised: June 26, 2020
Accepted: July 16, 2020

Corresponding Author:
Keun Ho Ryu
khryu@tdtu.edu.vn,
khryu@chungbuk.ac.kr

Publisher:
Chaoyang University of Technology
ISSN: 1727-2394 (Print)
ISSN: 1727-7841 (Online)

https://doi.org/10.6703/IJASE.202009_17(3).223

About 54% of Korean smokers want to quit smoking even smoking cessation law and negative impacts of health are publicized in Korea population. In 1986, the government implemented a smoking cessation policy concerning for smoking damage. The government has been launching a smoking cessation program for public health centers for strengthening anti-smoking policies since 1999. Smoking rate has been slightly decreased in terms of implementation of the government's expansion of tobacco control policies. However, the smoking rate in Korea is still higher than the world average 17.3% in 2015 (WHO, 2015; Lee and Seo, 2007). Especially, government and health care providers initiated to implement accessible resources in order to help quit smoking. According to a report released by the World Health Organization (WHO, 2017), the incremental asking for anti-smoking comes from the awareness, which was continuously addressed by the smoke free organization and committed health care requirement spread over society as well. And it effects to build smoking prevention policies to offer smoke free in society. Therefore, people should be able to breathe tobacco-smoke-free air and many countries have been realizing to decrease tobacco consumption through monitoring and implementing smoke-free ways for encouraging smokers to quit effectively. An important component of the smoking cessation program is the understanding of the factors and predicting success for quitting which is an effective way for public health benefit.

Previous research studies have highlighted the significant factors associated with smoking cessation. Kim (2014) focused on identifying significant predictors of successful smoking cessation using a large representative sample of the KNHANES dataset. Researchers estimated weight gain would be associated with early that post-cessation weight gain would predict relapse over and the influence of other variables, such as treatment condition, baseline nicotine dependence, and gender (Borrelli et al., 2001). The study by Kim (2012) evaluated smoking prevalence for Korean adults by gender, age group and the association between smoking and socio-demographic factors using the KNHANES dataset. This study concerned the high smoking prevalence among widowed or divorced women also it conducted with a cross-sectional analyze and using to estimate Rao-Scott Chi-square test, Crude odds ratio and confidence intervals in 95% for finding association and comparison of variables. The study (Charafeddine et al., 2017) estimated the association between health-related quality of life and smoking for each educational level and gender using linear and logistic multivariate regression models. Among women, however, daily smokers have shown the significantly lower health-related quality of life scores compared with never smokers, but only among females with a low and intermediate educational level. A majority of studies compared to estimate objectives and applied statistical methods such as chi-square test, logistic and multivariate regression models for finding the association between socio-demographic factors and success for smoking cessation. The regression analysis estimates statistical significant interactions among the dependent variable and one or more independent variables.

Nowadays classification technique plays an essential role in drive the medical decision rules effectively (Davagdorj et al., 2019; Davagdorj et al., 2020; Lin et al., 2004). Classification is supervised learning in which the predictor learns from the data input and the objective of a classification model is to predict the target class with the most accurate result. Real-world data mining problems involve learning classifiers from class imbalanced data, which means that one of the classes include a small number of values than other classes. Class imbalance problem mostly appears in applications such as remote sensing, credit scoring, fraud detection and especially medical diagnosis (Leichtle et al., 2017; Huang et al., 2006; Marqués et al., 2013; Sahin et al., 2013; Ganji et al., 2010; Babar and Ade, 2015). According to the study by Maciejewski and Stefanowski (2011), detecting rare events is a prediction problem, which is difficult to detect their infrequency and casualness. Class Imbalance problem constitutes a difficulty for most learning algorithms, which are biased toward learning and recognition of the majority classes, as well as, minority examples tend to be misclassified.

One of the basic strategies for addressing the class imbalance problem is data resampling (Zheng et al., 2016). Resampling techniques are used to improve class imbalance learning in order to alleviate the effect of the skewed class distribution in the learning process. Under-sampling and over-sampling techniques are most commonly used to rebalance the sample space for class imbalanced sample. Moreover, under-sampling is a technique to reduce the number of samples in the majority class, where the size of the majority class sample is reduced from the original datasets to balance the class distribution. However, a drawback of under-sampling is the loss of information. On the other hand, over-sampling techniques increase the number of minority class members in the training set. Duplication of random records is named its drawback because it can be a cause of overfitting.

This paper presents a SMOTE based decision support framework in order to solve the class imbalance problem in the smoking cessation program. Our proposed framework has three main steps: First, we generate the target data and determine significant features with multivariate analysis. Second, three variations of the SMOTE is applied in training set to generate synthetic records for balancing minority class with the majority class. The final step is performance evaluation; we will propose the best prediction model in the success of smoking cessation program. Finally, compare the prediction models to obtain the best model that measured by the accuracy, precision, recall and f-score in imbalanced and balanced data. This paper considers the following contributions:

- Provide the significantly important factors of smoking based on statistical bivariate analyze and multivariate regression analysis.
- SMOTE technique to handle imbalanced dataset, it also

resolves biased prediction of classifiers that may arise due to the class imbalance problem.

- Provide the best combination of over-sampling techniques and classification algorithms for constructing the prediction model for smoking cessation program.
- These findings can enhance the understanding of the significant factors and efficient prediction models for program implementation of smoking cessation and accompanying to concern public health.

The remainder of this paper is logically structured as follows: Section 2 describes the procedure of decision support framework. Section 3 presents the baseline characteristics of the dataset and experimental results. Section 4 includes the discussion of the findings. Finally, Section 5 concludes the paper.

2. MATERIALS AND METHODS

In this paper, we have proposed synthetic oversampling based decision support framework for the smoking cessation program, which divided into main three steps as illustrated in Fig. 1. The overall procedure of the experimental framework and method are described in this section.

In the first step, we perform preprocessing for the smoking cessation raw dataset in order to handle missing values and outliers as well as we select a subset of relevant features for use in model construction. Moreover, we will make the factor analyze among significant features to interpret important risk factors. In the second step, three variations of SMOTE technique are used to perform oversampling for the imbalanced dataset, which is expected to solve the class imbalance problem. In the third step, machine learning classification algorithms are compared and concerned with smoking cessation predictive modeling using training data which be formed by the imbalanced and balanced dataset. Finally, the prediction performances are measured by accuracy, precision, recall and F-score, respectively.

2.1 Step 1: Data preprocessing and factor analysis

Feature selection is an essential preprocessing step in data mining for finding an optimal subset of relevant features as well as improving performance for classifiers from the original dataset. Significant filter feature selection is used to eliminate redundant and irrelevant features. We used the chi-square test to find the significant features for the dependent variable or class label. The features are considered significant when the p-value is less or equal than 0.05. In case of higher than given threshold value of 0.05, we exclude those features, because the features are not separated in terms of dependent variables. After performing the feature selection, we also applied the multivariate logistic regression analysis to select risk factors for smoking cessation.

2.2 Step 2: Over-sampling - Synthetic Minority Oversampling Technique (SMOTE)

Resampling techniques are used to address imbalance learning in order to alleviate the effect of the skewed class distribution in the learning process. Over-sampling techniques increase the number of minority class members in the training set. Most importantly, the proposed over-

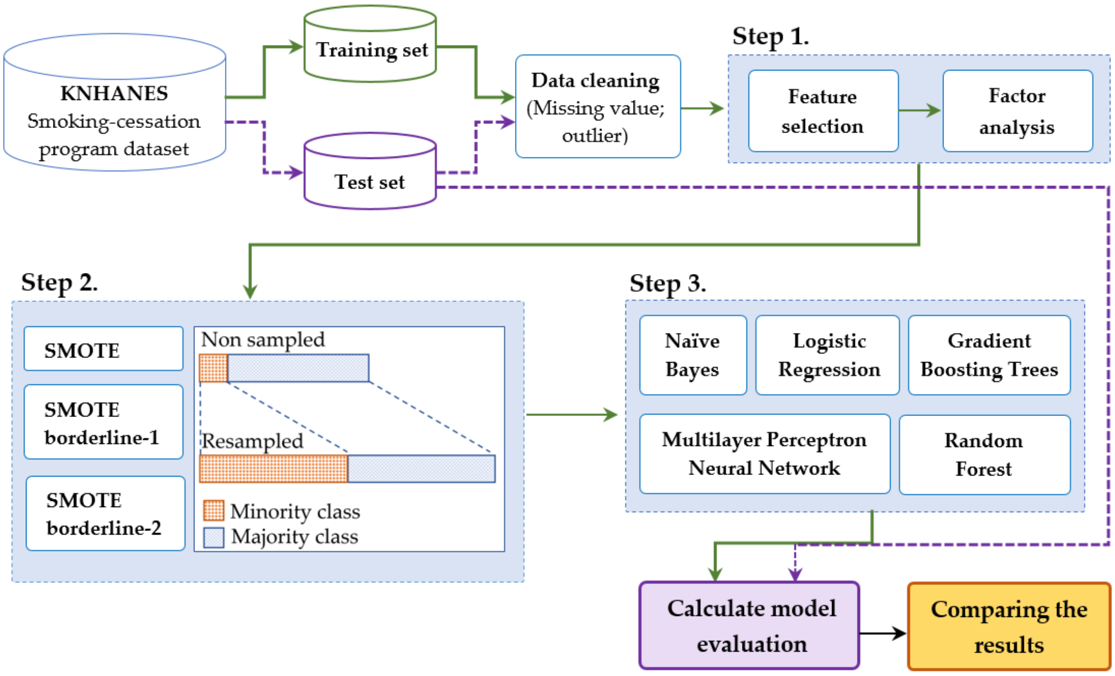

Fig. 1. Synthetic oversampling based decision support framework for smoking cessation program.

sampling approach called SMOTE in which the minority class is over-sampled by creating “synthetic” examples rather than by over-sampling with duplicated real data entries in recent year (Chawla et al., 2002).

SMOTE generates synthetic samples of minority class based on feature space instead of data space in randomly sampling methods. There are three different variations of SMOTE algorithms: SMOTE regular, SMOTE borderline-1 and SMOTE borderline-2 (Han et al., 2005). For the SMOTE regular, the underlying algorithm is designed to deal with continuous (or discrete) variables. Given an imbalanced dataset, the minority class is oversampled by taking each minority class instance and creating the number of synthetic samples along with the K nearest neighbors of the minority class. The parameter N that is number of synthetic samples generated per original minority instance, and the K related to nearest neighbors, which need to be predefined. Pseudo code of SMOTE regular is presented in Algorithm 1.

The SMOTE borderline, include borderline-1 and borderline-2, will classify each data point x_i to be (i) noise when all nearest neighbors come from a different class than the one of x_i, (ii) in danger (when at least half of the nearest neighbors are from a different class than x_i), x_i or (iii) safe when most of nearest neighbors are from the same class as x_i. The SMOTE borderline-1 and borderline-2 will use the data points in danger to generate new data points.

2.3 Step 3: Classification Algorithms

Naïve Bayes (NB) is statistical classifier that represents the relationship between the prior and posterior probability of random variables (Rish, 2001). Bayes probability analysis suggests that Bayes theorem can find the posterior probability from the prior probability. Bayes theorem is important when mathematically dealing with decision problems under uncertainty. This is useful when calculating the value of invisible intangible assets such as information. NB classifier examines the notion of conditional probability.

Algorithm 1. SMOTE algorithm

```
Input: Number of minority class samples T; Amount of SMOTE N%;
            Number of nearest neighbors k ;
Output: (N/100) * T // synthetic minority class samples
    if N < 100;
    // If N is less than 100%, randomize the minority class samples as only a random percent of them will be
  SMOTE.
            then Randomize the T minority class samples;
                T = (N/100) * T;
            N=100;
    endif;
    N = (int)(N/100); // The amount of SMOTE is assumed to be in integral multiples of 100.
    k = (int); // Number of nearest neighbors
    numattrs = (int); // Number of attributes
    Sample[ ][ ]; // array for original minority class samples
    newindex; // keeps a count of number of synthetic samples generated, initialized to 0
    Synthetic[ ][ ]; // array for synthetic samples
    // Compute k nearest neighbors for each minority class sample only
    for i ←1 to T;
            // Compute k nearest neighbors for i, and save the indices in the nnarray
            Populate(N, i, nnarray);
    end for Populate(N, i, nnarray); // Function to generate the synthetic samples
    while N≠0;
            // Choose a random number between 1 and k, call it nn
            // This step chooses one of the k nearest neighbors of i
            for attr ←1 to numattrs;
                Compute: dif = Sample[nnarray[nn]][attr] − Sample[i][attr];
                Compute: gap = random number between 0 and 1;
                Synthetic[newindex][attr] = Sample[i][attr] + gap * dif;
            endfor;
            newindex++;
            N = N – 1 ;
    endwhile;
    return; // End of Populate
    End of Pseudo-Code.
```

The conditional independence probability is calculated as:

$$P(X|Y=y)=\Pi_{t=1}^{d}P(X_i|Y=y) \qquad (1)$$

Where Y is class label and $X=\{X_1,X_2,\dots,X_d\}$ is the individual attribute set which is comprised of d attributes. For the conditional independence assumption, instead computing the class-conditional probability for every combination of X, it estimates the conditional probability of each X_i and given Y.

The NB classifier computes the posterior probability for each class Y. Mathematically formula can be written as:

$$P(X|Y)=\frac{P(Y)\Pi_{t=1}^{d}P(X_i|Y)}{P(X)} \qquad (2)$$

Logistic Regression (LR) is a statistical method for analyzing a dataset where the dependent variable is categorical (Menard, 2002). The goal of logistic regression predicts the probability of an outcome that only has two possible dichotomy values (successful quitter or unsuccessful quitters for smoking), which is limited to values between 0 and 1, from a set of independent variables. The logistic model is shown by:

$$F(x)=\frac{1}{1+e^{-x}} \qquad (3)$$

Odds ratio represents how many times the probability of success is higher than the probability of failure as show below:

$$odds\ ratio=\frac{p(y=1|x)}{1-p(y=1|x)} \qquad (4)$$

The confidence level indicates the probability that the confidence interval will contain the true odds ratio.

Multilayer Perceptron Neural Network (MLPNN) was inspired by the biological neural system and multilayer perceptron is the most typical type of neural networks (Basheer and Hajmeer, 2000). In the structure of the multilayer perceptron neural network, the neurons are split into typically three layers, which are called input, hidden and output. Each input multiplied by the corresponding weight parameter and these weighted sums is transferred to a hidden layer, and then produce the output layer. The network utilized types of the activation function (e.g., sigmoid or tanh) to produce the output. Finally, determine optimized weights of nodes by minimizing the error between actual and prediction target.

Random Forest (RF) is a class ensemble tree-based method which bagging to generate subsets of the entire training set to build multiple individual decision trees (Liaw and Wiener, 2002). Ensemble classifier aggregates the individual predictions to combine into a final prediction voting for the most popular class. This classification technique required the main two kinds of parameters such as a number of trees and number of attributes used to grow each tree. For instance, one popular advantage for using random forest over single decision tree classifier is reducing over-fitting of training data and get more accurate. The reason of it, we used random forest ensemble method for predicting success or unsuccessful reason for smokers as well.

Gradient Boosting Tree (GBT) is a widely used machine-learning algorithm due to its efficiency, accuracy and interpretability (Ke et al., 2017). Boosting methods are based on the idea that converting weak learners into strong learners. Gradient boosting produces a prediction model in the form of an ensemble of typically decision trees. This trains many models in a gradual, additive and sequential manner, and it uses the gradient in the loss function. The loss function is a measure indicating how well the model's coefficients can fit the underlying data.

2.4 Evaluation metrics

In order to perform the model comparison, accuracy, precision, recall and F-score evaluation metrics are used (Powers, 2011). These classification metrics are determined four value of confusion matrix, which contains the information about actual and predicted values in classification as shown in Table 1. True Positive (TP) means that the predicted value is positive and the actual value is positive. False Negative (FN) means that the predicted value is negative and the actual value is positive. False Positive (FP) means that the predicted value is positive and the actual value is negative. True Negative (TN) means that the predicted result is negative and the actual value is negative (Luque et al., 2019).

According to the confusion matrix, the evaluation metrics were defined as shown in Equation 5-8. Accuracy is how well the classifier can predict a positive case positive and negative case negative. Error rate is when the classifier predicted a positive case negative and the negative case positive. The accuracy can be computed as below:

$$Accuracy=\frac{TP+TN}{(TP+TN+FP+FN)} \qquad (5)$$

Precision is the fraction of relevant instances among the retrieved instances as seen in Equation 6:

$$Precision=\frac{TP}{TP+FP} \qquad (6)$$

For recall metric, the fraction of the total amount of relevant instances that were actually retrieved as shown in Equation 7.

$$Recall=\frac{TP}{TP+FN} \qquad (7)$$

The F score is obtained as the weighted harmonic mean of the test's precision and recall. Moreover, F-score is mostly suggested evaluation metric for imbalanced analysis. F-score can be criticized in particular circumstances due to its bias as an evaluation metric as shown in Equation 8:

$$F=2\cdot\frac{precision\ \cdot\ recall}{precision\ +\ recall} \qquad (8)$$

https://doi.org/10.6703/IJASE.202009_17(3).223

3. EXPERIMENT AND RESULT ANALYSIS

3.1 Data Preprocessing

The sample selection procedure is shown in Fig. 2. In total, 47,674 subjects are included in this study. Firstly, we exclude subjects aged less than 18 years old, because of the smoking-related survey collected from subjects aged equal or more than 18 years old. Then, we analyze the successful quitters with current smokers who had tried to quit but failed their attempt. The successful quitters are defined as those who reported that they had smoked at least 100 cigarettes in their life and not smoke currently about more than 12 consecutive months. The current smokers who had a recent attempt to quit but failed are defined as those who reported that they had smoked at least 100 cigarettes in their life, do smoke currently, and had stopped smoking for more than one day but failed during the past 12 months.

More importantly, we have to consider those successful and unsuccessful smoking quitters should be attending in at least one smoking cessation program. For instance, the dataset includes smoking cessation counseling, smoking cessation agent, public health center, cessation clinic, and doctor-prescribed medication. Therefore, we exclude 5,869 smokers who did not try to quit smoking in their life. Moreover, 6,661 number subjects are included in our study, because they attended at least one smoking cessation program. The 2,868 number of missing value and outliers are removed in order to evaluate the acceptable prediction performance. Standard deviation method for detecting outliers by using criteria value (0.05) lognormal distribution is used. Finally, we select totally 3,793 subjects. In this study, we consider the unsuccessful smokers of 2,982 who attended in smoking cessation program but failed and successful quitters are 811 who can quit smoking after smoking cessation program.

3.2 Feature selection and factor analysis

According to the related studies of the smoking cessation (Kim, 2014; Davagdorj et al., 2019), we selected initial features from the KNHANES dataset. Using the chi-square test, we compared features between smoking successful and unsuccessful quitters' groups. The features are considered significant when the p-value is less or equal to 0.05. In case of higher than given threshold value of 0.05, we excluded those features because these features are not separated in terms of dependent features. From the result of chi-square test, "Gender" (p = 0.658), "Household Income" (p = 0.836), "Age of smoking initiation" (p = 0.173) and "Average sleep time per day" (p = 0.645) features have been excluded because p-value of threshold is higher than 0.05. On the contrary, "Age (year)", "Education", "Occupation", "Marital status", "Subjective health status", "Frequency of alcohol consumption in recent 1 year", "Body mass index", "Stress", "Secondhand smoke in the workplace", "Daily smokers at home" and "Attendance in smoking prevention or smoking cessation education"

Table 1. Confusion matrix for two-class problem.

Predicted class Actual class	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

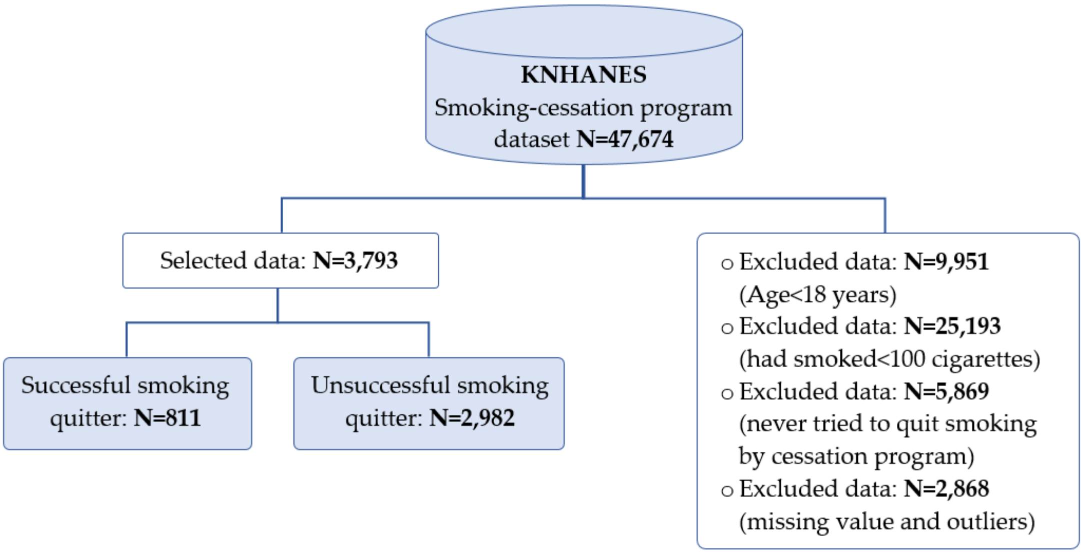
KNHANES
Smoking-cessation program dataset N=47,674
Selected data: N=3,793
Successful smoking quitter: N=811
Unsuccessful smoking quitter: N=2,982
○ Excluded data: N=9,951 (Age<18 years)
○ Excluded data: N=25,193 (had smoked<100 cigarettes)
○ Excluded data: N=5,869 (never tried to quit smoking by cessation program)
○ Excluded data: N=2,868 (missing value and outliers)

Fig. 2. Sample selection procedure.

Table 2. The result of factor analysis using multivariate logistic regression.

Features	OR	CI 95%
Age (year)		
Less than 35	ref	
36-45	0.230	0.174~0.303
46-55	0.332	0.267~0.414
More than 56	0.506	0.415~0.618
Education		
Below to elementary school graduate	ref	
Middle school graduate	0.630	0.477~0.832
High school graduate	0.729	0.556~0.955
College graduate or higher	0.845	0.704~1.014
Marital status		
Married	ref	
Single and others	1.065	0.807~1.406
Occupation		
Managers, experts or related workers	ref	
Office workers	4.834	2.680~8.720
Service or seller	4.687	2.578~8.523
Agriculture, forester or fisher	4.138	2.291~7.474
Function, device or machine assembly worker; farmers	2.937	1.613~5.348
Labor workers	3.725	2.075~6.685
Housewife, students or other	3.693	2.026~6.733
Subjective health status		
Very good	ref	
Good	1.496	0.760~2.944
Normal	1.378	0.743~2.556
Bad	1.379	0.746~2.547
Very bad	1.062	0.566~1.995
Frequency of alcohol consumption in recent 1 year		
Not drink at all in the last year	ref	
Less than once a month	1.595	1.211~2.102
About once a month	1.265	0.958~1.671
2-4 times a month	1.690	1.263~2.262
About 2-3 times a week	1.300	1.037~1.631
4 or more times a week	1.048	0.833~1.319
Body mass index		
Lower weight (13.0-19.0)	ref	
Normal weight (20.0-25.0)	0.758	0.619~0.927
Over or severe obese weight (26.0-40.0)	0.926	0.793~1.080
Level of perceived stress		
Very high	ref	
Moderate	0.646	0.424~0.984
Low	0.840	0.673~1.098
Rarely	0.991	0.804~1.222
Indirect secondhand smoking exposure in the workplace		
Yes	ref	
No	1.559	1.226~1.984
No working place	1.576	1.238~2.007
Daily smokers in the home		
Yes	ref	
No	0.431	0.325~0.573
Smoking prevention or smoking cessation education		
Yes	ref	
No	0.431	0.327~0.568

features have been indicated significant features due to threshold of p-value is less than 0.05.

After the significant feature selection, multivariate analysis is performed by the multivariate logistic regression. In multivariate logistic regression analysis, the odds ratio (OR) and 95% confidence interval (CI) of features for smoking status are reported in Table 2.

The successful smoking abstinence is significantly related to increasing age. More than 56 years old aged individuals (OR = 0.506, 95% CI = 0.415~0.618) are more likely to quit than less than 35 years old people. Therefore, a college graduate or higher educated individuals (OR = 0.845, 95% CI = 0.704~1.014) likely to be quit than individuals who graduated below or elementary school. The occupation is found to be one of the significant indicators of smoking cessation due to the program. A function, device or machine assembly worker and farmers are 2.937 times (OR = 2.937, 95% CI = 1.613~5.348) more likely to be quit than managers, experts or related workers. Among the highest frequency of alcohol use as a four or more times a week, it reduced the likelihood of smoking quit by (OR = 1.048, 95% CI = 0.833~1.319). Furthermore, the lowest level of subjective health status increased the likelihood of unsuccessful smoking cessation by (OR = 1.062, 95% CI = 0.566~1.995). Also, increasing the odds of body mass index indicator by normal weight by (OR = 0.758, 95% CI = 0.619~0.927) and over or severe obese weight by (OR = 0.926, 95% CI = 0.793~1.080) are more associated with smoking successful cessation. Individuals who had rare stress are 0.991 times more likely (OR = 0.991, 95% CI = 0.804~1.222), to have a positive relationship with smoking quitters compared with those who have not. Moreover, individuals who have a daily smoker in the home are (OR = 0.431, 95% CI = 0.325~0.573), and those have an attended in smoking cessation education in order to prevention by 0.431 times more likely reduced the likelihood of smoking cessation.

3.3 Oversampling

From the data-preprocessing result, 811 successful quitters and 2,982 unsuccessful quitters (that means they have failed their attempt even try to quit smoking) have resulted. Totally, 3,793 individuals have resulted from data pre-processing step. In the next step, we used SMOTE regular, borderline-1 and borderline-2 techniques for creating a synthetic minor class. Moreover, nearest neighbor is an important parameter for SMOTE over-sampling technique. This closest nearest neighbor numbers can decide the artificial values which create in minority class during the training process. The created dataset is considered to address the class imbalance problem. Closest neighbor K numbers of SMOTE regular, SMOTE borderline-1, SMOTE borderline-2 are adjusted as 3, 5 and 7. Therefore, these techniques can achieve the desired ratio between the majority and minority classes. We set the 1:1 ratio between binary class label for successful and unsuccessful smoking quitters. This over-sampled set was then used for training the benchmark classifiers.

3.4 Evaluation results

The primary goal of the comparisons is to evaluate the effectiveness of the combination with various over-sampling techniques with NB, LR, MLPNN, RF and GBT models due to predict the success of smoking quit after attend in a smoking cessation program. To obtain a reliable predictive measurement, we have used 5-fold stratified cross-validation. In 5-fold cross-validation, data is randomly partitioned into 5-fold. From it, four folds are used as the training set and one set is used as the testing test. For evaluating the prediction model in case of class imbalance problem, values of the minor class have not occurred in some partition. Prevent this issue; stratified cross-validation is used efficiently because the partitions are selected so that the mean response value is approximately equal in all the partitions.

According to the evaluation metrics, Tables 3 through 6 present the results of the benchmark models without synthetic oversampling technique and three kinds of SMOTE techniques, respectively. The highest result for each metric is in bold.

Prediction results of the different prediction models on imbalanced data sample are seen in Table 3. As evaluation performances of 811:2982 imbalanced data sample, RF classifier outperforms the best precision score of 0.7587, recall score of 0.8214, F-score of 0.7751 and accuracy of 0.8214, respectively.

In the case of Table 4, the dataset has been oversampled using SMOTE-regular technique. As a result, it is evidently seen that in terms of precision, NB scores the highest performance by 0.8313, followed by MLPNN with a score of 0.8122 in 811:2982 sample data. On the contrary, the RF classifier outperforms the best results in metrics for recall by 0.8402 and F-score by 0.8202, respectively.

As shown in Table 5, RF classifier achieves the best accuracy of 0.8514, recall of 0.8311 and F-score of 0.8203 when dataset is balanced by the SMOTE borderline-1. Contrary, NB classifier reaches the best precision score of 0.8304, significantly. Moreover, SMOTE borderline-1 with NB classifier performs the acceptable accuracy of 0.8063, but it gives the worst recall score of 0.4992 and F-score of 0.5819, respectively.

For the Table 6, SMOTE borderline-2 with NB classifier performs the best precision score of 0.8310. On the contrary, SMOTE borderline-2 with RF gives the highest accuracy of 0.8267, recall of 0.8067 and f-score of 0.8061, respectively. In terms of the F-score, second best model is MLPNN, which scores 0.7540, following it, GBT with 0.7320 and LR with 0.6950. The worst F-score of 0.5847 is performed by the NB classifier.

According the each evaluation metric, we compare the performance of the prediction models in imbalanced and balanced data is depicted in Fig. 3 through 6. We have used three variations SMOTE techniques such as SMOTE regular, borderline-1 and borderline-2 in imbalanced dataset.

As shown in Fig. 3, GBT classifier outperforms the worst accuracy of 0.6753 in imbalanced data. On the contrary, RF classifier reaches the highest accuracy of 0.876 when SMOTE regular is used. In terms of the precision score, the lowest 0.5521 score is performed by MLPNN classifier. SMOTE regular with NB classifier reaches the best precision score of 0.8313 as well as three variations of SMOTE can significantly improve the precision score in each classifier as illustrated in Fig. 4. According to the recall score in Fig. 5, RF classifier reaches the highest score of 0.8214 compare with other classifiers in imbalanced dataset as well as this classifier combines with SMOTE regular slightly improves to 0.8402. Therefore, SMOTE borderline-1 with NB classifier performs the lowest recall of 0.4992.

Table 3. Comparison results of prediction models on imbalanced data.

Classifier Metric	NB	LR	MLPNN	RF	GBT
Accuracy	0.7988	0.8174	0.6973	**0.8214**	0.6753
Precision	0.5810	0.7523	0.5521	**0.7587**	0.6418
Recall	0.6299	0.5059	0.5746	**0.8214**	0.5099
F-score	0.5786	0.4685	0.5524	**0.7751**	0.4796

Table 4. Comparison results of prediction models on balanced data using SMOTE regular.

Classifier Metric	NB	LR	MLPNN	RF	GBT
Accuracy	0.8138	0.8032	0.8198	**0.8760**	0.7082
Precision	**0.8313**	0.7977	0.8122	0.8038	0.8063
Recall	0.5038	0.6332	0.5921	**0.8402**	0.6682
F-score	0.5862	0.6961	0.6655	**0.8202**	0.7223

Table 5. Comparison results of prediction models on balanced data using SMOTE borderline-1.

Classifier Metric	NB	LR	MLPNN	RF	GBT
Accuracy	0.8063	0.7923	0.8011	**0.8514**	0.7712
Precision	**0.8304**	0.7977	0.8136	0.8108	0.8065
Recall	0.4992	0.6332	0.7169	**0.8311**	0.7002
F-score	0.5819	0.6961	0.7567	**0.8203**	0.7440

Table 6. Comparison results of prediction models on balanced data using SMOTE borderline-2.

Classifier Metric	NB	LR	MLPNN	RF	GBT
Accuracy	0.8143	0.7527	0.7423	**0.8267**	0.7549
Precision	**0.8310**	0.7974	0.8146	0.8056	0.7979
Recall	0.5023	0.6317	0.7123	**0.8067**	0.6849
F-score	0.5847	0.6950	0.7540	**0.8061**	0.7320

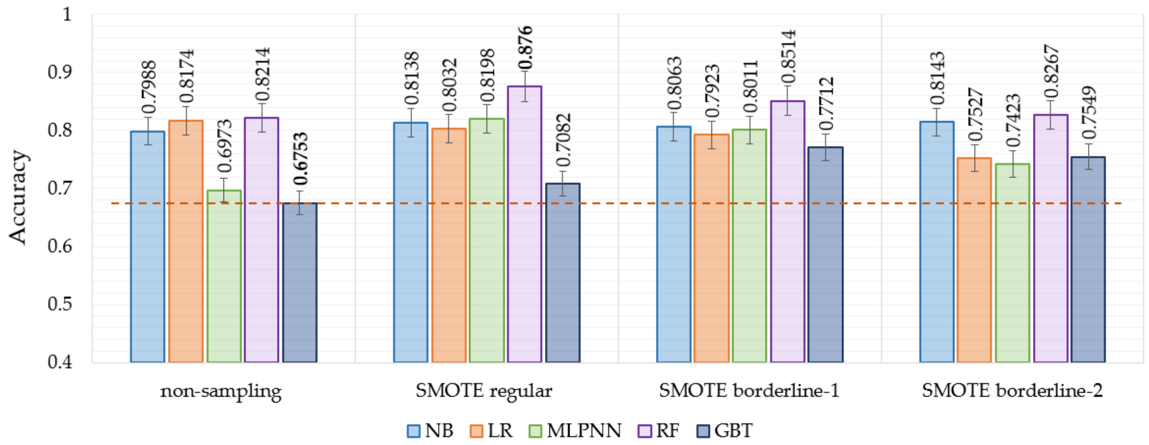

Fig. 3. Accuracy of the prediction models among imbalanced and balanced data.

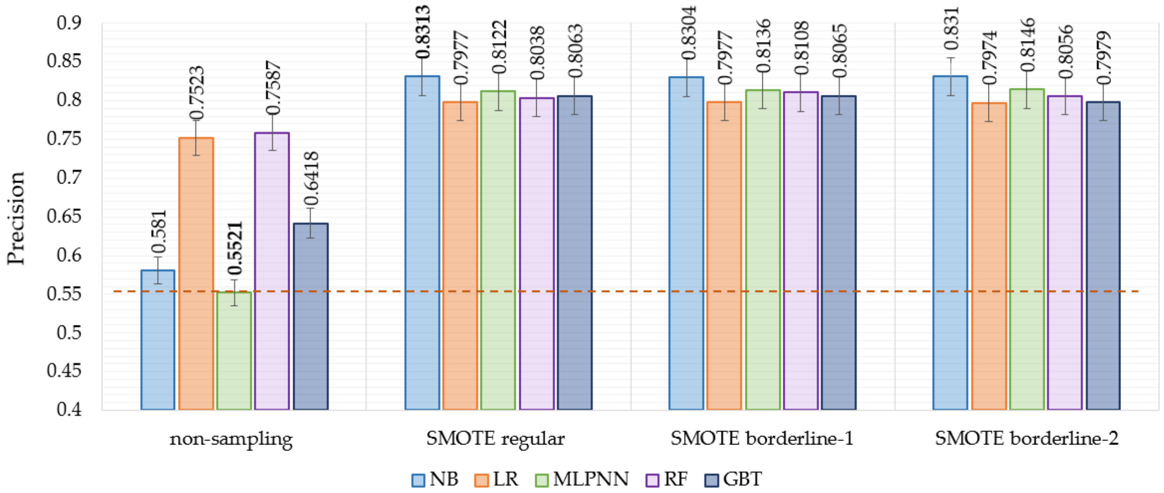

Fig. 4. Precision of the prediction models among imbalanced and balanced data.

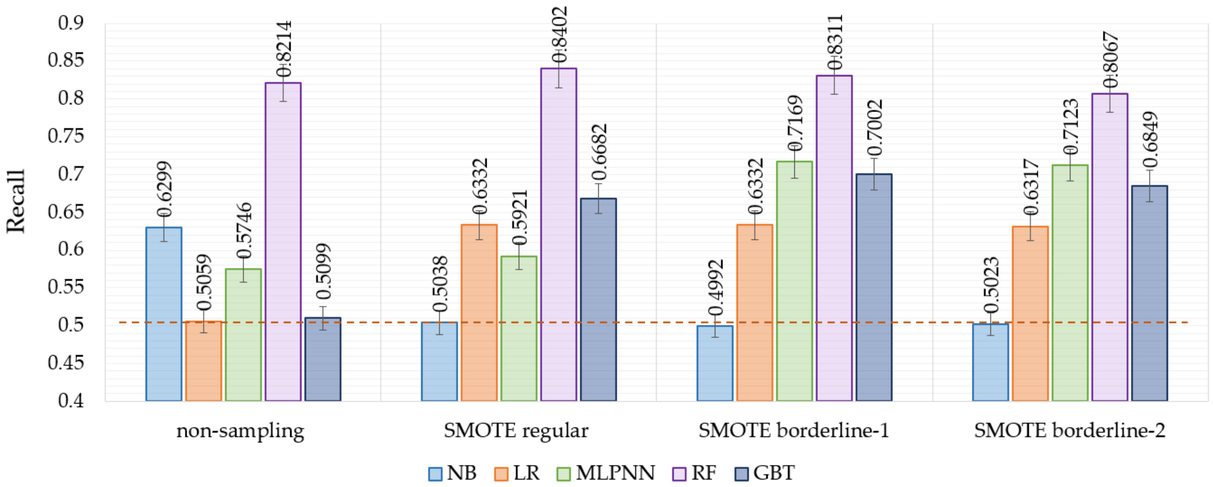

Fig. 5. Recall of the prediction models among imbalanced and balanced data.

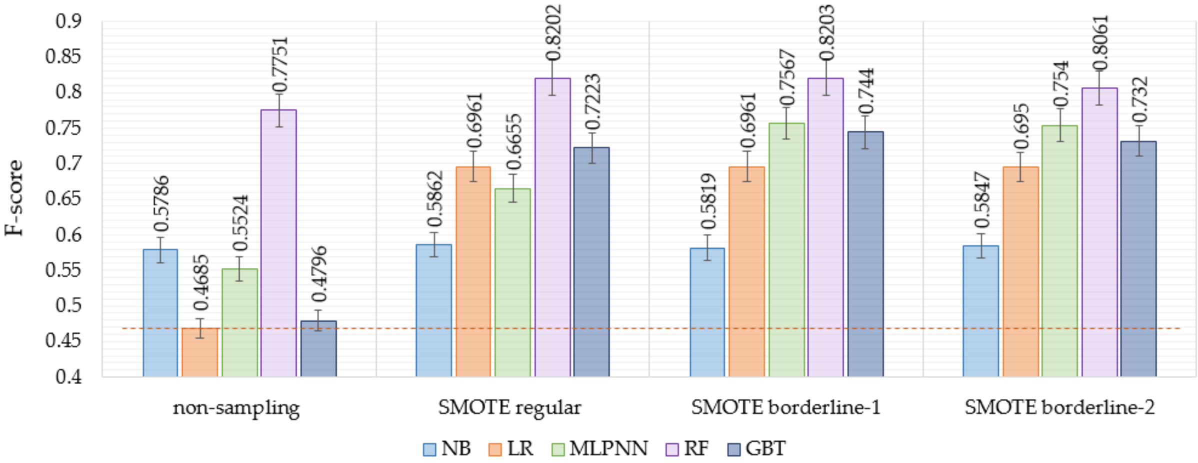

Fig. 6. F-score of the prediction models among imbalanced and balanced data.

In Fig. 6, RF classifier reaches the highest score of 0.7751 in imbalanced dataset. Moreover, combination of the SMOTE regular, borderline-1 and borderline-2 with RF classifier outperform the highest results of the 0.8202, 0.8203 and 0.8061, respectively. In terms of the F-score, synthetic oversampling technique can improve the prediction results as well as SMOTE regular and SMOTE borderline-1 methods reach the computable higher results.

4. DISCUSSIONS

Smoking is the main cause of the various chronic diseases, cancers and early death among not only smokers but also second-hand smokers. Smoking cessation is an important task to reduce the risks of cancers such as lung, larynx, esophagus, and pancreas. In this paper, we have presented an efficient decision support framework for predicting the smoking cessation program using real-world KNHANES dataset. Our experimental framework consists of main three steps: the first step eliminates the unnecessary instances and variables applying chi-square test and multivariate logistic regression analysis; the second step utilizes three variations of SMOTE such as regular, borderline-1 and borderline-2 to in order to solve the class imbalance problem; and machine learning classifiers constructs the prediction models in the third step. Finally, compare the prediction models to obtain the best model that measured by the accuracy, precision, recall and F-score in imbalanced and balanced data.

In the first step, the individuals who are former or current smokers have been chosen while the sample selection procedure. Essentially, one of the important criteria was that they should be tried to quit by at least one kind of smoking cessation program in their life. According to the classification outcome, other individuals eliminated from our target dataset. In terms of the chi-square test and multivariate logistic regression analysis, we choose the most significant variables. Age, education and frequent alcohol use are important predictors in smoking cessation success. Furthermore, the lowest level of subjective health status has increased the likelihood of unsuccessful smoking cessation. In the second step, we used three variations of SMOTE to create synthetic instances in minor class; and five different machine learning classifiers construct the prediction models in the third step. To obtain a reliable predictive measurement, we have used 5-fold stratified cross-validation in terms of the imbalanced class distribution.

For determining the effect of the three variations of SMOTE, we compared the prediction results of machine learning classifiers among imbalanced and balanced data. The experimental result presents that RF classifier outperforms the best precision score of 0.7587, recall score of 0.8214, F-score of 0.7751 and accuracy of 0.8214 in imbalanced dataset. In terms of class imbalance problem, SMOTE regular, borderline-1 and borderline-2 were used to enhance the prediction performance. As a result, SMOTE with classifiers significantly improves the results, respectively. Moreover, NB classifiers reached the highest precision score when it used SMOTE regular. On the contrary, SMOTE with RF achieved the best scores of recall, F-score and accuracy, as well as SMOTE based RF model determined the best prediction model for smoking cessation program.

5. CONCLUSION

In this paper, we presented the synthetic oversampling technique based decision support framework in order to predict the success of smoking cessation program, which obtains smoking quitter and non-quitter classes. In particular, we analyzed the publicly available KNHANES dataset. In the real world, analyzing class imbalance is one of the quite challenge to learn model, thus we have applied efficient three variations of SMOTE in our model training process. For constructing the prediction model for the success of smoking cessation program, we adopted five different machine-learning classifiers. Our experimental results showed that NB classifier performed the best precision score of 0.8313, and RF classifier achieved the highest accuracy of 0.876, recall of 0.8402 and F-score of 0.8202 when we use the SMOTE regular in imbalanced dataset. Moreover, SMOTE significantly improves the performance of the smoking cessation prediction model. This finding has gone some way towards enhancing our understanding of prediction in this area. In further work, the proposed framework can be extended by deep learning methods which address the problems of class imbalance and human-readability. Empirically, more comparison findings would be a substantial benefit for public health.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2017 R1A2B4010826), and (No. 2019K2A9A2A06020672) and (No.2020R1A2B5B02001717).

REFERENCES

Babar, V., Ade, R. 2015. A novel approach for handling imbalanced data in medical diagnosis using under sampling technique. In Communications on Applied Electronics (CAE), Foundation of Computer Science FCS.

Basheer, I.A., Hajmeer, M. 2000. Artificial neural networks: fundamentals, computing, design, and application. Journal of microbiological methods, 43, 1, 3–31.

Borrelli, B., Spring, B., Niaura, R., Hitsman, B., Papandonatos, G. 2001. Influences of gender and weight gain on short-term relapse to smoking in a cessation trial. Journal of Consulting and Clinical Psychology, 69, 3, 511.

Charafeddine, R., Demarest, S., Cleemput, I., Van Oyen, H., Devleesschauwer, B. 2017. Gender and educational differences in the association between smoking and health-related quality of life in Belgium. Preventive medicine, 105, 280–286.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321–357.

Davagdorj, K., Lee, J.S., Park, K.H., Ryu, K.H. 2019, October. A machine-learning approach for predicting success in smoking cessation intervention. In 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), IEEE, 1–6.

Davagdorj, K., Lee, J.S., Pham, V.H., Ryu, K.H. 2020. A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention. Applied Sciences, 10, 9, 3307.

Davagdorj, K., Yu, S.H., Kim, S.Y., Huy, P.V e., Park, J.H., Ryu, K.H. 2019. Prediction of 6 months smoking cessation program among women in Korea. International journal of machine learning and computing, 9, 1, 83–90.

Ganji, M.F., Abadeh, M.S., Hedayati, M., Bakhtiari, N. 2010, November. Fuzzy classifcation of imbalanced data sets for medical diagnosis. In 2010 17th Iranian Conference of Biomedical Engineering (ICBME), 1–5, IEEE.

Han, H., Wang, W.Y., Mao, B.H. 2005. August. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing, 878–887, Springer, Berlin, Heidelberg.

Huang, Y.M., Hung, C.M., Jiau, H.C. 2006. Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Analysis: Real World Applications, 7, 4, 720–747.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 3146–3154.

Kim, S. 2012. Smoking prevalence and the association between smoking and sociodemographic factors using the Korea National Health and Nutrition Examination Survey Data, 2008 to 2010. Tobacco Use Insights, 5, TUI–S9841.

Kim, Y.J. 2014. Predictors for successful smoking cessation in Korean adults. Asian nursing research, 8, 1, 1–7.

Lee, E.S., Seo, H.G. 2007. The factors associated with successful smoking cessation in Korea. Journal of the Korean Academy of Family Medicine, 28, 1, 39–44.

Leichtle, T., Geiß, C., Lakes, T., Taubenböck, H. 2017. Class imbalance in unsupervised change detection–a diagnostic analysis from urban remote sensing. International journal of applied earth observation and geoinformation, 60, 83–98.

Liaw, A., Wiener, M. 2002. Classification and regression by randomForest. R news, 2, 3, 18–22.

Lin, C.-J., Peng, C.-C., Lee, C.-Y. 2004. Prediction of RNA polymerase binding sites using purine-pyrimidine encoding and hybrid learning methods. International Journal of Applied Science and Engineering, 2, 2, 177-188.

Luque, A., Carrasco, A., Martín, A., de las Heras, A. 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216–231.

Maciejewski, T., Stefanowski, J. 2011. April. Local neighbourhood extension of SMOTE for mining imbalanced data. In 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 104–111, IEEE.

Marqués, A.I., García, V., Sánchez, J.S. 2013. On the suitability of resampling techniques for the class imbalance problem in credit scoring. Journal of the Operational Research Society, 64, 7, 1060–1070.

Menard, S. 2002. Applied logistic regression analysis, Series: Quantitative Applications in the Social Sciences, 106, Sage publications, International Educational and Professional Publisher.

Powers, D.M. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2, 1, 37–63.

Rish, I. 2001. August. An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence, IBM New York, 3, 22, 41–46.

Sahin, Y., Bulkan, S., Duman, E. 2013. A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, 40, 15, 5916–5923.

Song, Y.M., Sung, J., Cho, H.J. 2008. Reduction and cessation of cigarette smoking and risk of cancer: a cohort study of Korean men. Journal of clinical oncology, 26, 31, 5101–5106.

World Health Organization and Research for International Tobacco Control, 2008. WHO report on the global tobacco epidemic, 2008: the MPOWER package. World Health Organization.

World Health Organization, 2015. WHO report on the global tobacco epidemic 2015: raising taxes on tobacco. World Health Organization.

World Health Organization, 2017. WHO report on the global tobacco epidemic, 2017: monitoring tobacco use and prevention policies. World Health Organization.

Zheng, Z., Cai, Y., Li, Y. 2016. Oversampling method for imbalanced classification. Computing and Informatics, 34, 5, 1017–1037.