applied sciences

MDPI

Article

A Comparative Analysis of Machine Learning Methods for Class Imbalance in a Smoking Cessation Intervention

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

[1] Database and Bioinformatics Laboratory, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, Korea; suri@dblab.chungbuk.ac.kr (K.D.); richard@dblab.chungbuk.ac.kr (J.S.L.)

[2] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam; phamvanhuy@tdtu.edu.vn

[3] Department of Computer Science, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, Korea

* Correspondence: khryu@tdtu.edu.vn or khryu@chungbuk.ac.kr; Tel.: +82-10-4930-1500

† These authors are contributed equally to this work.

Received: 26 March 2020; Accepted: 6 May 2020; Published: 9 May 2020

Abstract: Smoking is one of the major public health issues, which has a significant impact on premature death. In recent years, numerous decision support systems have been developed to deal with smoking cessation based on machine learning methods. However, the inevitable class imbalance is considered a major challenge in deploying such systems. In this paper, we study an empirical comparison of machine learning techniques to deal with the class imbalance problem in the prediction of smoking cessation intervention among the Korean population. For the class imbalance problem, the objective of this paper is to improve the prediction performance based on the utilization of synthetic oversampling techniques, which we called the synthetic minority over-sampling technique (SMOTE) and an adaptive synthetic (ADASYN). This has been achieved by the experimental design, which comprises three components. First, the selection of the best representative features is performed in two phases: the lasso method and multicollinearity analysis. Second, generate the newly balanced data utilizing SMOTE and ADASYN technique. Third, machine learning classifiers are applied to construct the prediction models among all subjects and each gender. In order to justify the effectiveness of the prediction models, the f-score, type I error, type II error, balanced accuracy and geometric mean indices are used. Comprehensive analysis demonstrates that Gradient Boosting Trees (GBT), Random Forest (RF) and multilayer perceptron neural network (MLP) classifiers achieved the best performances in all subjects and each gender when SMOTE and ADASYN were utilized. The SMOTE with GBT and RF models also provide feature importance scores that enhance the interpretability of the decision-support system. In addition, it is proven that the presented synthetic oversampling techniques with machine learning models outperformed baseline models in smoking cessation prediction.

Keywords: smoking; class imbalance; synthetic oversampling; machine learning; decision making; feature importance

1. Introduction

Cigarette smoking is the most avoidable risk factor for premature death. The World Health Organization (WHO) estimated that the tobacco epidemic is currently responsible for the death of more than 8 million people across the world each year. Although, the number of deaths may increase to approximately 10 million by the year 2020 [1,2]. The WHO report noticed that half of the world's smokers are alive today. However, they are more likely to die prematurely due to tobacco-related disease if they continue to smoke. Indeed, tobacco contains high addictive substance namely nicotine, which is the leading cause of smoking dependence. Nicotine addiction can be strongly associated with a variety of psychiatric diagnoses, cancers, stroke, and chronic disorders. On the other hand, smoking cessation reduces the chance of cancer risk such as lung, larynx, esophagus, and pancreas among smokers [3–7]. Most tobacco users make multiple quit attempts of smoking over their lifetime and regret having started. Nevertheless, the immensely addictive nature of nicotine makes it hard to quit smoking dependence for most smokers without effective interventions [8,9]. Due to its complexity, many countries have been realizing to decrease tobacco consumption in the care of planning and implementing smoke-free ways. Thus, there remains a need to construct an efficient decision support system, thereby assisting in smoking cessation intervention for determining whether to quit smoking or not.

It is widely known that machine learning techniques have been efficiently used in the healthcare area, including disease and risk prediction [10–13]. However, decision-making responses suffer from the class imbalance problem. The class imbalance problem has received much attention from researchers in the fields of medical [14], fraud detection [15] and bankruptcy prediction [16]. To deal with this problem, advanced data-driven and machine learning techniques have been developed constantly. The sampling techniques are investigated to rebalance an imbalanced dataset to alleviate the effect of the skewed class distribution. Broadly, sampling techniques can be classified into two groups, such as under-sampling and over-sampling [17]. Under-sampling discards the samples from the majority class to make it equal to the minor class. A drawback of under-sampling is the loss of information. On the other hand, over-sampling creates the samples in minority class to make it equal to the majority class. Moreover, it has a drawback, which is associated with duplicated random records, which can be a cause of overfitting. Existing studies have exhibited that the synthetic minority over-sampling technique (SMOTE) and adaptive synthetic (ADASYN) over-sampling techniques are used rather than downsizing dataset [18–20].

Therefore, we conduct an empirical comparison based on machine learning methods to predict the success of smoking cessation, having real-world smoking cessation interventions dataset collected from the Korea National Health and Nutrition Examination Survey (KNHANES). Accordingly, we aim to solve the class imbalance problem where utilizing synthetic oversampling techniques, namely SMOTE and ADASYN. Generally, our experimental design is composed of the following components: First, a lasso and multicollinearity—based feature selection approach is addressed to eliminate irrelevant features. Next, SMOTE and ADASYN techniques are applied to create new synthetic data for balancing the classes among all subject, men and women. Finally, widely used logistic regression (LR), multilayer perceptron (MLP), deep MLP, Random Forest (RF), Gradient Boosting Trees (GBT), k nearest neighbors (KNN) and support vector machine (SVM) classifiers are utilized to build the prediction models in imbalanced and balanced datasets. Comparison analyses determine the proper setting of hyper-parameters for minimizing the classification error. The non-parametric Friedman test is applied to ensure the differences between performances of prediction models. Moreover, SMOTE with GBT and RF models are provided feature importance scores in order to enhance the interpretability of the decision support system in smoking cessation. Overall, the principal contributions of this study are as follows:

- Provide a comparative analysis of machine learning methods based on experimental design to determine their predictive performances when used to predict smoking cessation intervention among the Korean population.
- Experimental results prove that the integration of synthetic oversampling techniques with machine learning classifiers enhance the evaluation performance among all subjects and each gender.
- Important features are defined in terms of lasso and multicollinearity analysis, which can be a substantial benefit for understanding the causes of smoking risk factors.

- SMOTE with GBT and RF models are provided feature importance scores to enhance the model interpretability.
- Empirically, practical models have gone some way towards solving implication for practice class imbalanced problem, and these findings would be significantly important for the public health impact has arisen from the quit smoking.

The remainder of this paper is organized as follows: Section 2 introduces the literature review of smoking cessation. Section 3 presents our proposed experimental design, including the building procedure and experimental configuration in detail. In Section 4, the dataset and overall experimental results are provided. Section 5 presents the discussion of this study. Finally, the study is concluded in Section 6.

2. Literature Review

The area of smoking cessation becomes a widely researched topic. Numerous studies [21–28] have focused on the prediction model of smoking cessation using statistical methods such as bivariate analysis and multivariate logistic regression (LR). In the study of Monso, E. et al., [21], highly related features were assessed by age, sex, and housing conditions among European smokers who attended in intervention programs voluntarily. In a study by Kim, Y. J [22], the frequency of alcohol consumption and trying numerous quitting methods were inversely related to smoking cessation among the Korean population. Charafeddine et al. [23] analyzed the association between health-related quality of life and smoking among individuals varied by gender and educational level. Their findings highlighted that health-related quality of life scores are not associated with smoking among men, but significantly associated with smoking among women with lower education. In another study, Lee, S. et al. [24] developed a prediction model for future smoking intention among Korean adolescent using the cross-sectional school-based Korea Global Youth Tobacco Survey in order identify high risk group exposed to future smoking. They utilized essentially five determinants past smoking experience: parents' smoking status; friends' smoking status; ownership of a product with a cigarette brand logo; and intentions of smoking from close friends' cigarette offered in their study.

Another study [25] focused on the effect of inpatient counseling follow-up and predictors of successful smoking cessation analysis. They found that smoking quit success was highly positively associated with smoking cessation counseling and its frequency of counseling number during hospitalization. Foulds, J et al. [26] studied risk factors associated with successful quitting at a free tobacco treatment clinic at 4-week and 6-month follow-up. Low socioeconomic status and high nicotine dependence were estimated negative effect on the success of smoking cessation outcomes. Therefore, Smit, E. S et al. [27] investigated the predictors of successful and unsuccessful quit attempts among smokers motivated to quit within 6 months. According to the findings from these researchers, self-efficacy appeared to play a role in the main factor for predicting quit attempts and their success. Another study by Blok et al. [28] addressed the impact of individual potential predictors, with a special focus on respiratory and cardiovascular disease. They studied a longitudinal study in a large cohort of smokers. In their study, smoking cessation and relapse were highly affected by smokers in household members and friends.

Nowadays, a limited amount of studies [29–31] demonstrated that machine learning classifiers, such as decision tree, SVM, and MLP have the potential to substitute statistical methods in constructing the prediction models of smoking cessation. Coughlin et al. [29] presented the prediction model of smoking quit outcomes to improve the success of evidence-based tobacco use treatment and current methods. The significant features were selected by generalized estimating equations followed by using decision tree classifier to identify the smoking quit prediction. Another study proposed by Poynton et al. [30] applied a neural network for the classification of smoking cessation status among current and former smokers. They identified a significant feature subset in their examined adult data sample derived from the National Health Interview Survey in 2000. More recently, one of our previous studies Davagdorj et al. [31] carried out a study to compare the machine learning classifiers, thereby suggesting that the response of smoking quit success after 6 months of attendance in the smoking cessation program. To achieve the goal, we examined real-world data on smoking cessation programs among women in our previous study. The datasets were obtained from Chungbuk Tobacco Control Center of Chungbuk National University College of Medicine in South Korea, from 2015 to 2017. Accordingly, the statistically significant features were used as input to train predictive models using a different machine learning classifier.

Based on the previous findings on smoking cessation area, LR analysis is still considered the baseline model used to identify statistically significant and associated factors, based on the observed characteristics of the target group. Most of the existing studies tended to ignore comprehensive prediction models and its hyper-parameter tuning. In addition, machine-learning-based studies in smoking were not highlighted the model interpretability.

3. Materials and Methods

Our stated experimental design for constructing the prediction model of smoking cessation intervention can be divided into three components. In the first component, we preprocess the smoking cessation raw dataset to eliminate missing values and outliers, further electing a subset of appropriate features for use in model construction. In the second component, SMOTE and ADASYN techniques derive the balanced data. This component is expected to resolve biased classifier performance in an imbalanced binary class problem. Eventually, machine learning classification algorithms are used to build predictive models to tackle the success of smoking cessation based on imbalanced and balanced datasets among all subjects, men and women. Comparison findings would be to reveal a suitable combination of synthetic oversampling techniques and machine learning classifiers in this domain. The flowchart of the proposed experimental design is depicted in Figure 1, among the remainder of this section detailing these components.

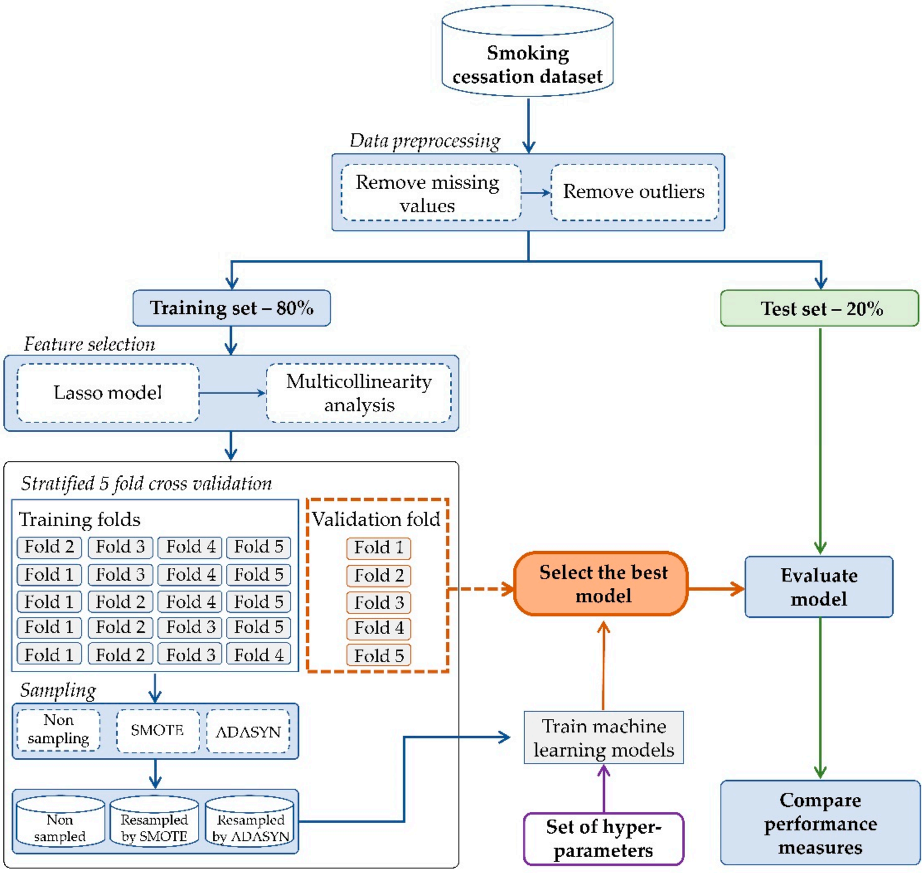

Figure 1. Experimental design for smoking cessation intervention.

3.1. Data Preprocessing

Typical real-world datasets have useful information, but it is not in a form appropriate for the required data mining procedure. Thus, data pre-processing and feature selection are applied to a given set of raw data. First, this component considers removing all missing values and outliers. One of the relevant problems in data quality is a missing value, which occurs when data values are not stored in the feature. The absence of data values can significantly affect the representativeness of the sample and statistical power. Therefore, the estimation of those values is important to build prediction models accurately. Outlier refers to the observation that has abnormal distance from other values on tails of the distribution. Therefore, the lost data and outliers can cause bias in the estimation of parameters and complicate the analysis of the study.

Second, the lasso method [32] and multicollinearity-analysis-based feature selection approach [33] excludes redundant and irrelevant features in an effort to reduce the complexity of the training model. Lasso eliminates the number the features in terms of their significance of the dependent variable. Furthermore, a rational decision is taken to execute the multicollinearity analysis in order to demonstrate the correlation between independent features. The value of the variance inflation factor is used to verify multicollinearity in regression analysis. In essence, this component takes the complete set of features and loops through all of them applying the appropriate test. After that, if deemed significant, the features are kept for the next component. Otherwise, such features are eliminated.

3.2. Sampling Techniques

To address the imbalance problem, we utilize SMOTE [34] and ADASYN [35] techniques, which are the commonly used benchmark oversampling algorithms. The SMOTE regular technique was designed by Chawla et al. in 2002, where the minority class concerns synthetic minority class samples that are homogeneously distributed around the original positive instances that lie close together. The synthetic oversampling technique can produce artificial data instead of general random replacement of actual samples. Depending on the oversampling rate, a certain number of samples from the k nearest neighbors are randomly chosen. The synthetic samples are generated as follows: estimate the difference between an instance and its nearest neighbor, which occur in minor class. Later, a random number among 0 and 1 multiplies this difference. Finally, add it to the feature vector that corresponds to the new synthetic instances of the minority class.

ADASYN is known as the pioneer of developing algorithms that are derived by synthetic techniques, including SMOTE. The overall process for ADASYN adaptively generates more synthetic data of minority class instances according to their weighted distributions using nearest neighbors. The main idea of this technique is that it can adaptively generate synthetic data for a minor class to decrease the bias, which occurred from the data imbalance problem. Therefore, it is using weighted distributions of the different minority class in terms of their level of difficulty in learning, thus it improves the learning performance.

3.3. Machine Learning Classifiers

This study compares the performances of the various numbers of classifiers within the context of smoking cessation intervention, thereby assessing different imbalanced and balanced dataset. A brief explanation of each of the machine learning classifiers, such as LR [36], MLP [37], deep MLP [38], RF [39], GBT [40], k nearest neighbors (KNN) [41] and SVM [42], used in this paper is given below.

LR is an extensively used statistical method for solving classification and regression problems. Therefore, we will be focusing on the binary response of whether a smoker can quit or not in this paper. Thus, LR can predict the probability of an outcome that only has two possible dichotomy values, which is limited to values between 0 and 1, from a set of independent features.

MLP is the most typical type of neural network application using back propagation for training. Neural networks are inspired are composed of nodes. MLPs consist of at least three layers, such as input, hidden and output. Nodes in neighboring layers are interconnected, but nodes in the same layer are not. Each connection between neurons is multiplied by the corresponding weight during training. Finally, the output of hidden nodes is estimated by applying an activation function and output layer makes decisions.

Ensemble methods: ensemble methods can be categorized into two parts, such as parallel and sequential. RF is a parallel structured ensemble tree-based method that utilizes bagging to aggregate multiple decision tree classifiers. Each tree of the RF is trained on bootstrap samples of the training sets using randomly selected features in the tree generation process; after this, each tree votes for the most popular class. GBT is an effective method in sequential ensemble learning of either regression or classification tree models. It consists of a gradient boosting and regression decision tree, which uses an ensemble of classification, and regression trees by the base learner. Generally, GBT combines a series of decision trees sequentially; each tries to correct its predecessor in order to build the strongest one.

KNN belongs to supervised learning, and instances are classified to the class most frequently occurring amongst the neighbors that measured by the distance function. In a sense, the number of k-neighbors and the number of distance metrics are important parameters in the KNN classification phase. In the case of the smallest k, the classifier may tend to overfit because of the noise in the training set. Conversely, too large values of k and the classifier may misclassify because the nearest neighbors include instances that even stayed far away from their neighborhood.

Support Vector Machine: SVM is a powerful method for building classifiers. It finds a decision boundary known as the hyperplane. Here, optimal hyperplane separates instances correctly into each class, and the margin width between the optimal hyperplane and the training instances is maximized to fit the data. The function is used to map the instances.

3.4. Experimental Setup

For evaluating the prediction models, we split the data into 80% for the training set and 20% for the evaluation set. The 5-fold stratified cross-validation procedure [43] is applied to the training set. In the procedure of 5-fold stratified cross-validation, the dataset is randomly partitioned into five parts, 4 folds of the training set are used to train classification models, and the remaining 1 fold is used to validate the model. In the presence of class imbalance, stratified cross-validation is a commonly used method because folds are made by an equal number of samples for each class to preserve optimally balanced distributions. This procedure can end up providing enough representatives of minority and majority classes in each fold. Principally, several machine learning classifiers need to set up a few parameters is shown in Table 1.

Table 1. Range of parameters for classification models.

Classifier	Parameter	Range
RF	Number of estimators Criterion	250, 500, 750, 1000, 1250, 1500 gini; entropy
GBT	Learning rate Number of estimators	0.15, 0.1, 0.05, 0.01, 0.005, 0.001 250, 500, 750, 1000, 1250, 1500
KNN	Number of k range Weight options	3, 6, 9 uniform, distance

Grid search has been applied to hyper-parameter tuning to find the optimal ones. For the MLP models, we use the one and three hidden layers with 5 nodes. These models are optimized by Adam, and we set the learning rate is 0.001 with "sigmoid" activation function. In the RF classifier, the number of estimators has selected 250, 500, 750, 1000, 1250, 1500 trees, and the quality of split-measured criteria is employed "gini" for the Gini impurity and "entropy" for the information gain, respectively. The GBT classifier is to use learning rates and the number of estimators, which are calculated as the optimal combinations to materialize better performance. The hyper-parameters of the KNN classifier are weights and number of neighbors. The weights are either set up to "uniform", where all points in each neighborhood are weighted the same, or "distance", where closer points are more heavily weighted toward the decision. The setting of the neighbor numbers refers to how many neighboring points are to fall inside of one group.

In our study, a rank-based non-parametric Friedman test [44] is used to compare the various machine learning classifiers. The Friedman test is computed as follows:

$$\chi_F^2 = \frac{12D}{K(K+1)}\left[\sum_{k=1}^{K} AvR_j^2 - \frac{K(K+1)^2}{4}\right]$$

where D is the number of datasets and K is the number of classification algorithms. R_j is the rank of the jth classification algorithm on a dataset. If chi-square distribution χ_F^2 with $(d-1)$ degrees of freedom is larger than a critical value, the null hypothesis is rejected.

The f-score, type I error, type II error, balanced accuracy and geometric mean evaluation metrics [45,46] are utilized to compare the classifier's performance and select the best model. F-score is a metric that considers both precision and sensitivity. Precision is calculated as the number of correct positive results divided by the total number of results that the model identified as positive. For smoking cessation intervention, type I error appears when successful smoking quitters are misclassified as failed, whereas type II error occurs when smokers who failed in during smoking cessation intervention are misclassified as a successful quitter. In more detail, type I error and type II error are defined as the type I error = 1 − specificity and type II error = 1 − sensitivity, respectively. Most algorithms are designed to maximize the standard accuracy due to class imbalance. Instead of this, balanced accuracy is suggested to evaluate the prediction models. It is defined as the arithmetic mean of sensitivity and specificity. The geometric mean is efficiently utilized as the performance metric that is computed as a square root of the sensitivity multiplied by specificity.

All experiments were performed on PC with 3.20 GHz, Intel Core i5, and 8GB RAM using the Microsoft Windows 10 operating system. Thus, the first component of a study has been executed by Statistical Package for the Social Sciences (SPSS) and the experiment of the comparison analysis has been carried out by Python with the open libraries such as Pandas, Numpy, Matplotlib, scikit-learn, SciPy and Imbalanced-learn [47–52] and so on.

4. Experiment and Result Analysis

According to Figure 1, the proposed experimental design requires the execution of the number of components: (I) feature selection, (II) oversampling following by (III) comparison between synthetic oversampling techniques and machine learning classifiers for prediction models of the smoking cessation intervention. Moreover, non-parametric Friedman's test is applied to ensure the differences between the results of the prediction models are significant. In addition, the best representative features were provided, which were used to build the best ranked predictive models.

4.1. Subjects and Dataset

The experimental dataset is taken from the Korea National Health and Nutrition Examination Survey (KNHANES), which is conducted by the Korea Centers for Disease Control and Prevention (KCDC). The KNHANES dataset contains a health examination for various numbers of diseases, health interview, and nutrition surveys. In this way, some considerable necessary features were surveyed from a minor of the population with around only one year. Therefore, the dataset contains features that have a large number of missing values and outliers. We have analyzed the dataset collected from 2009 to 2017 among the Korean population, after acquiring permission from the KCDC (http://knhanes.cdc.go.kr).

Regarding the classification target, it needs to generate the target population for eliminating irrelevant features and instances. Additionally, only adults aged 18 years or older who were considered for smoking-related surveys. Current and former smokers were defined as those who have smoked at least 100 cigarettes in their life. On the contrary, individuals who have never smoked or have smoked less than 100 cigarettes in their life were considered as non-smokers. Initially, the total number of subjects was 75,292, constituting the total number of the collected data. Similar to the literature [22,23,31], we selected 22 features. Subjects were eliminated from the current study based on the following exclusion criteria:

- We excluded 53,850 subjects who never smoked.
- We excluded 9271 subjects who never tried to quit smoking cigarettes in their life by smoking cessation methods.

The smoking cessation methods of the KNHANES dataset consist of interventions such as smoking quit agent, smoking quit clinic of the public health center, doctor prescribed medication, and no smoking guides by telephone, internet and various kind of counseling, which were used to determine class labels. Then, 8479 subjects were excluded due to missing value and outliers, which were stored in given initial features. The outliers and extreme values were removed based on the interquartile range. The exclusion process of missing value and outlier produced the 3692 subjects with 22 features. According to the classification target, we use the "smoking status of adult" feature that has three groups such as smoking", "sometimes smoking" and "quit smoking". In this study, "smoking" and "sometimes smoking" are considered as unsuccessful smoking quitters, whereas "quit smoking" is defined as a successful quitter, which means a person who used to smoke but had not smoked in the past 1 year.

Descriptive characteristics of the former and current smokers are provided in Table 2. Consequently, 951 (25.7%) of the smokers can quit smoking habit successfully after attending at least one smoking cessation intervention voluntarily. This target class is recognized by imbalanced distribution. In comparison to socio-demographic features, we found evidence that approximately 90% of subjects were men among former and current smokers. Men were more likely to smoke than women; thus, 856 (25.9%) of men and 95 (24.4%) of women were resulted by successfully smoking quitters in our study. For the age group, (<35 years) subjects were less likely to quit smoking than other age groups. Taken together, 14.09% of <35 years, 21.90% of 36–45 years, 26.30% of 46–55 years and 40.89% of >56 years had successfully quit smoking. Among the body mass index, lower weights (13.0–19.0) of subjects were markedly less likely to quit smoking by about 7% than other groups. Among the smoking cessation rates by the urban and rural population, urban areas had two times higher successful quit rates than rural. Therefore, the subjects starting smoking in age groups 27.01% of <18 years, 23.40% of 19–24 years and 28.05% of >25 years were estimated to quit smoking, respectively. Thus, subjects who started to smoke in the age group of 19–24 years are less likely to quit smoking than other age groups.

Table 2. Basic characteristics of the subjects for smoking status.

Features	Former Smoker (%)	Current Smoker (%)
Gender		
Male	856 (23.2)	2447 (66.3)
Female	95 (2.6)	294 (8.0)
Age (year)		
Less than 35	129 (3.5)	786 (21.3)
36–45	223 (6.0)	795 (21.5)
46–55	217 (5.9)	608 (16.5)
More than 56	382 (10.3)	552 (15.0)
Body mass index		
Lower weight (13.0–19.0)	169 (4.6)	643 (17.4)
Normal weight (20.0–25.0)	483 (13.1)	1296 (35.1)
Over or severe obese weight (26.0–40.0)	299 (8.1)	802 (21.7)

Table 2. *Cont.*

Features	Former Smoker (%)	Current Smoker (%)
Household income		
Low	120 (3.3)	299 (8.1)
Low-middle	222 (6.0)	703 (19.0)
Middle-high	288 (7.8)	881 (23.9)
High	321 (8.7)	858 (23.2)
Residence area		
Urban	869 (23.5)	2237 (60.6)
Rural	82 (2.2)	504 (13.7)
Education		
Below to elementary school graduate	161 (4.4)	312 (8.5)
Middle school graduate	113 (3.1)	300 (8.1)
High school graduate	337 (9.1)	1092 (29.6)
College graduate or higher	340 (9.2)	1037 (38.1)
Employment condition		
Full time	1162 (31.5)	1444 (39.1)
Part time	500 (13.5)	442 (12.0)
Others	55 (1.5)	89 (2.4)
Occupation		
Managers, experts or related workers	192 (5.2)	520 (14.1)
Office workers	135 (3.7)	412 (11.2)
Service or seller	146 (4.0)	523 (14.2)
Agriculture, forester or fisher	136 (3.7)	233 (6.3)
Function, device or machine assembly worker; farmers	217 (5.9)	662 (17.9)
Labor workers	112 (3.0)	279 (7.6)
Housewife, students or other	13 (0.4)	112 (3.0)
Marital status		
Married	854 (23.1)	2245 (60.8)
Single and others	97 (2.6)	496 (13.4)
Subjective health status		
Very good	58 (1.6)	120 (3.3)
Good	327 (8.9)	903 (24.5)
Normal	443 (12.0)	1313 (35.6)
Bad	110 (3.0)	384 (10.4)
Very bad	13 (3.0)	21 (0.6)
Exercise		
Yes	24 (0.7)	494 (13.4)
No	927 (25.1)	2247 (60.9)
Frequency of alcohol consumption in recent 1 year		
Not drink at all in the last year	118 (3.2)	200 (5.4)
Less than once a month	105 (2.8)	287 (7.8)
About once a month	93 (2.5)	200 (5.4)
2–4 times a month	270 (7.3)	825 (22.3)
About 2–3 times a week	231 (6.3)	829 (22.5)
4 or more times a week	134 (3.6)	400 (10.8)
Hypertension		
Yes	464 (12.6)	844 (22.9)
No	487 (13.2)	1897 (51.4)

Table 2. *Cont.*

Features	Former Smoker (%)	Current Smoker (%)
Diabetes		
Yes	878 (23.8)	2111 (57.2)
No	73 (2.0)	630 (17.1)
Asthma		
Yes	139 (3.8)	285 (7.7)
No	812 (22.0)	2456 (66.5)
Average sleep time per day		
Less than 6 h	60 (1.6)	165 (4.5)
7–8 h	1129 (30.6)	1447 (39.2)
More than 9 h	383 (10.4)	508 (13.8)
Stress level		
Very high	30 (0.8)	148 (4.0)
Moderate	197 (5.3)	713 (19.3)
Low	587 (15.9)	1576 (42.7)
Rarely	137 (3.7)	304 (8.2)
Long term depression (more than two weeks)		
Yes	87 (2.4)	272 (7.4)
No	864 (23.4)	2469 (66.9)
Age of smoking initiation (years)		
Less than 18	516 (14.0)	1394 (37.8)
19–24	327 (8.9)	1070 (29.0)
More than 25	108 (2.9)	277 (7.5)
Secondhand smoke in the workplace		
Yes	468 (12.7)	1178 (31.9)
No	394 (10.7)	1146 (31.0)
No working place	89 (2.4)	417 (11.3)
Daily smokers at home		
Yes	49 (1.3)	320 (8.7)
No	902	2741 (74.2)
Attendance in smoking prevention or smoking cessation education		
Yes	55 (1.5)	296 (8.0)
No	896 (24.3)	2445 (66.2)

4.2. Feature Selection

The feature selection component consists of two phases, such as the lasso method and multicollinearity analysis. It is well known that the lasso method helps to increase the prediction of the model by removing irrelevant features that are not related to target classes. The lasso feature selection method is identified irrelevant features as "Residence area", "Employment condition", "Average sleep time per day", "Stress level" and "Long term depression (more than two weeks)" that eliminated by assigning them a coefficient equal to zero. On the contrary, the remaining 17 of the 22 features estimated along with non-zero coefficients, which were selected by sufficiently representative features.

We directly verify to check the collinearity between selected features using multicollinearity in regression analysis after eliminating the non-significant features, which are passed in the lasso method. Figure 2 illustrates the result of the Variance Inflation Factor (VIF) value for all independent features. Generally, it is suspected that multicollinearity will present if the VIF lies between 5 and 10. If the VIF value is greater than those values, it investigates a high correlation among features that remains problematic. As can be seen, none of the features have detected the presence of multicollinearity issue in the multiple linear regression models. In total, 17 features and 3682 subjects remained in the end of this component and are is used as inputs to the next components.

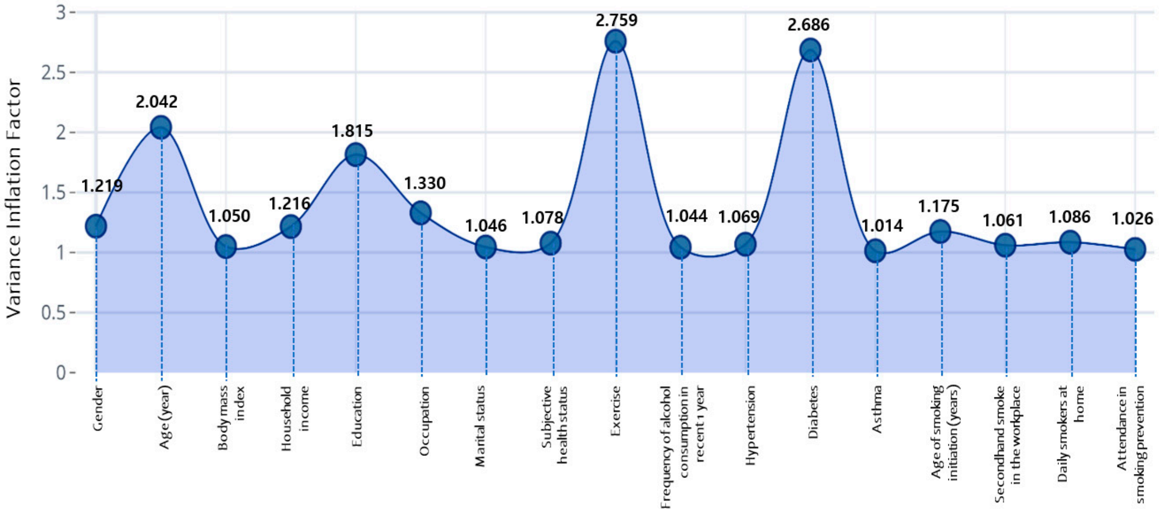

Figure 2. Results of the variance inflation factor analysis on the selected features.

4.3. Sampling

The experimental dataset includes 3692 subjects, which are recognized with the effect of the skewed class distribution with 951 successful (minor class) and 2741 unsuccessful (major class) smoking quitters. SMOTE and ADASYN oversampling techniques can achieve the desired ratio between the majority and minority classes. In our study, we configure the same between binary class labels within 1:1 ratios through the model training process, where k is set to 5 for SMOTE. Imbalanced and balanced datasets are used for training the prediction models in the third component.

4.4. Results and Comparison

In this section, we summarize the overall comparison results that were attained over various machine learning methods. Table 3 shows the results of the classifiers on imbalanced and balanced data among all subjects and highest performance of evaluation metrics are marked in bold.

As shown in Table 3, the KNN classifier performed the worst type II error of 0.2490 and F-score of 0.7604. On the contrary, the MLP classifier obtained the best type II error of 0.0500 and SVM achieved the highest F-score of 0.8343. As can be seen, it is hard to build the prediction model with a high geometric mean in terms of the highest type I error in imbalanced data; moreover, this error indicates that successful smoking quitters are misclassified as failed during the intervention. Dealing with the type I error issue under the class imbalance problem, we have applied the SMOTE and ADASYN oversampling techniques. The comparison result shows that the SMOTE with GBT classifier achieved the highest F-score of 0.8888, a balanced accuracy of 0.7928, a geometric mean of 0.7868 and a low type I error of 0.3044 when the learning rate was 0.005 and the number of the tree was 750. The lowest type II errors of 0.0405 and 0.0440 were achieved by the SMOTE with deep MLP, MLP classifiers, respectively. With regard to the evaluation metrics, SMOTE with GBT, MLP classifiers reached the highest results in all subjects.

In this study, our experimental dataset includes 3303 (89.46%) men and 389 (10.54%) women, whereas the former smokers' target includes 856 (23.20%) men and 95 (2.60%) women, and the current smokers' target includes 2447 (66.30%) men and 294 (8.00%) women. It is clearly shown that men have a tendency to use tobacco at higher rates than women; moreover, it is important to construct the prediction models for each gender. The evaluation results of prediction models among men and women are summarized in Tables 4 and 5.

Table 3. Evaluation results of the prediction models among all subjects.

Sampling	Model	F-Score	Type I Error	Type II Error	Balanced Accuracy	Geometric Mean
Non-sampling	LR	0.8016	0.9779	0.1027	0.4597	0.1408
	MLP	0.8048	0.9673	0.0500	0.4914	0.1763
	Deep MLP	0.7913	0.9701	0.0681	0.4809	0.1669
	RF	0.8148	0.8652	0.1763	0.4793	0.3332
	GBT	0.8179	0.9867	0.1600	0.4267	0.1057
	KNN	0.7604	0.7636	0.2490	0.4937	0.4214
	SVM	0.8343	0.9833	0.1040	0.4564	0.1223
SMOTE	LR	0.8171	0.6852	0.1500	0.5824	0.5173
	MLP	0.8660	0.5125	0.0440	0.7218	0.6827
	Deep MLP	0.8635	0.5456	0.0405	0.7069	0.6603
	RF	0.8571	0.3074	0.1500	0.7713	0.7673
	GBT	0.8888	0.3044	0.1100	0.7928	0.7868
	KNN	0.8056	0.3137	0.2264	0.7300	0.7286
	SVM	0.8410	0.4344	0.1300	0.7178	0.7015
ADASYN	LR	0.7498	0.6574	0.2045	0.5691	0.5221
	MLP	0.7690	0.5052	0.0705	0.7122	0.6782
	Deep MLP	0.7663	0.5311	0.0673	0.7008	0.6613
	RF	0.8227	0.4198	0.1433	0.7185	0.7050
	GBT	0.8339	0.3304	0.1141	0.7778	0.7702
	KNN	0.7894	0.3507	0.1900	0.7297	0.7252
	SVM	0.8187	0.4074	0.1550	0.7188	0.7076

Table 4. Evaluation results of the prediction models among men.

Sampling	Model	F-Score	Type I Error	Type II Error	Balanced Accuracy	Geometric Mean
Non-sampling	LR	0.4754	0.1971	0.6579	0.5725	0.5241
	MLP	0.4499	0.1496	0.6842	0.5831	0.5182
	Deep MLP	0.4522	0.1684	0.6814	0.5751	0.5147
	RF	0.7983	0.5000	0.2818	0.6091	0.5992
	GBT	0.8046	0.5055	0.2727	0.6109	0.5997
	KNN	0.2614	0.0474	0.8455	0.5536	0.3836
	SVM	0.2350	0.0365	0.8636	0.5500	0.3625
SMOTE	LR	0.8312	0.3827	0.2042	0.7066	0.7009
	MLP	0.7461	0.3698	0.3132	0.6585	0.6579
	Deep MLP	0.7305	0.3735	0.3340	0.6463	0.6459
	RF	0.8952	0.2405	0.1364	0.8116	0.8099
	GBT	0.7804	0.1970	0.3282	0.7374	0.7345
	KNN	0.7849	0.2307	0.2909	0.7392	0.7386
	SVM	0.7336	0.3523	0.3655	0.6411	0.6411
ADASYN	LR	0.8342	0.4237	0.2028	0.6868	0.6778
	MLP	0.7607	0.3984	0.3605	0.6206	0.6203
	Deep MLP	0.7527	0.4188	0.3619	0.6097	0.6090
	RF	0.8926	0.2627	0.1373	0.8000	0.7975
	GBT	0.7600	0.1558	0.3223	0.7610	0.7564
	KNN	0.7709	0.1842	0.3086	0.7536	0.7510
	SVM	0.7209	0.3263	0.3586	0.6576	0.6574

Table 5. Evaluation results of the prediction models among women.

Sampling	Model	F-Score	Type I Error	Type II Error	Balanced Accuracy	Geometric Mean
Non-sampling	LR	0.8554	0.9756	0.0159	0.5043	0.1550
	MLP	0.8551	0.9823	0.0167	0.5005	0.1319
	Deep MLP	0.8736	0.9824	0.0358	0.4909	0.1303
	RF	0.8130	0.9676	0.1169	0.4578	0.1692
	GBT	0.7874	0.9472	0.1525	0.4502	0.2115
	KNN	0.8154	0.9474	0.1017	0.4755	0.2174
	SVM	0.8299	0.9356	0.0807	0.4919	0.2433
SMOTE	LR	0.9047	0.5263	0.0339	0.7199	0.6765
	MLP	0.8348	0.4211	0.1864	0.6963	0.6863
	Deep MLP	0.8202	0.4595	0.2067	0.6669	0.6548
	RF	0.8065	0.7895	0.1525	0.5290	0.4224
	GBT	0.7627	0.7368	0.2373	0.5130	0.4480
	KNN	0.6275	0.5789	0.4576	0.4818	0.4779
	SVM	0.7833	0.7368	0.2034	0.5299	0.4579
ADASYN	LR	0.9134	0.5263	0.0169	0.7284	0.6824
	MLP	0.9000	0.3684	0.0847	0.7735	0.7603
	Deep MLP	0.8671	0.3903	0.1356	0.7371	0.7260
	RF	0.8160	0.7895	0.1356	0.5375	0.4266
	GBT	0.8130	0.7368	0.1525	0.5554	0.4723
	KNN	0.6214	0.6316	0.4576	0.4554	0.4470
	SVM	0.7934	0.7368	0.1864	0.5384	0.4628

For the data sample of men, the best model was distinguished by SMOTE with an RF classifier in terms of the f-score, type II error, balanced accuracy and geometric mean, which reached 0.8952, 0.1364, 0.8116 and 0.8099 when the number of trees is 250 and the criterion is entropy parameters. The SVM classifier performed the better type I error of 0.0365, but it performed the lowest geometric mean of 0.3625 due to worst type II error of 0.8636. Following this, the second best performances were yielded by the ADASYN with the RF classifier when the parameters for the number of trees was 250 and the criterion is gini, as shown in Table 4.

As seen in Table 5, the prediction models of women were hardly affected by the imbalanced distribution because the highest type I errors were reached by the baseline classifiers. In terms of type I errors, classifiers performed the lowest balanced accuracy and geometric mean metrics. However, ADASYN with MLP classifier exhibited better balanced accuracy of 0.7735, the geometric mean of 0.7603 and lowest type I error of 0.3684, significantly. In addition, the best f-score of 0.9134 was achieved by the ADASYN with LR.

Comparing to the evaluation performances are shown in Tables 3–5, certainly type I error issue is the biggest challenge in imbalanced data when it occurs our minority class for successful smoking quitters are misclassified as failed after smoking cessation intervention. Correspondingly, type II error occurs when current smokers are misclassified as successful quitters. Our comparison results revealed significant improvements in prediction models among all subjects and each gender when utilizing the SMOTE and ADASYN oversampling techniques. It is shown that geometric mean suits are one of the proper metrics for ascertaining between critical type I and type II errors. Therefore, concerning the geometric mean, Figures 3–5 illustrate the boxplot of the prediction models among all subjects, men and women.

As depicted in Figure 3, the GBT classifier presented the worst score in non-sampling data, but the SMOTE and ADASYN-based GBT classifier displayed significantly better results among all samples. For the data sample of men, RF and GBT ensemble models reached the best performance when it served the synthetic oversampling techniques, as illustrated in Figure 4. By way of contrast, the worst performances were performed by KNN and SVM single classifiers among men. Furthermore, ADASYN with the MLP classifier obtained the best prediction model, followed by ADASYN with deep MLP and LR classifiers presented the high scores in a data sample of women, as shown in Figure 5. Similar to the other samples, single LR, MLP, deep MLP and RF classifiers illustrated the lowest performance due to severe imbalance ratio and small sample data of women. To summarize, single models were inferior to the synthetic oversampling-based models over each sample.

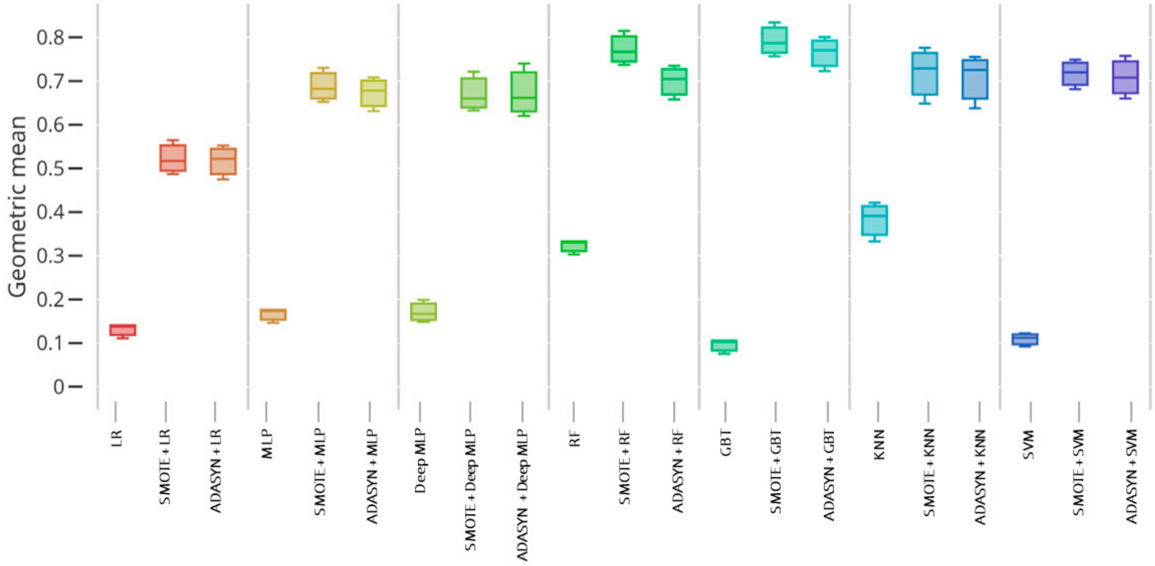

Figure 3. Boxplot of the geometric mean over prediction models among all subjects.

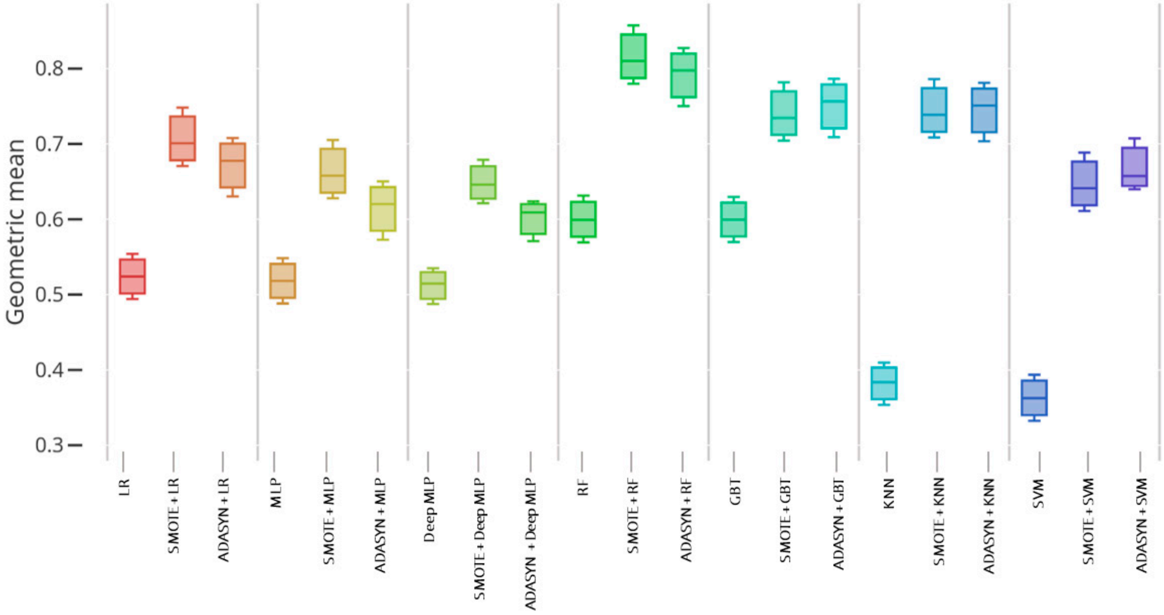

Figure 4. Boxplot of the geometric mean over prediction models among men.

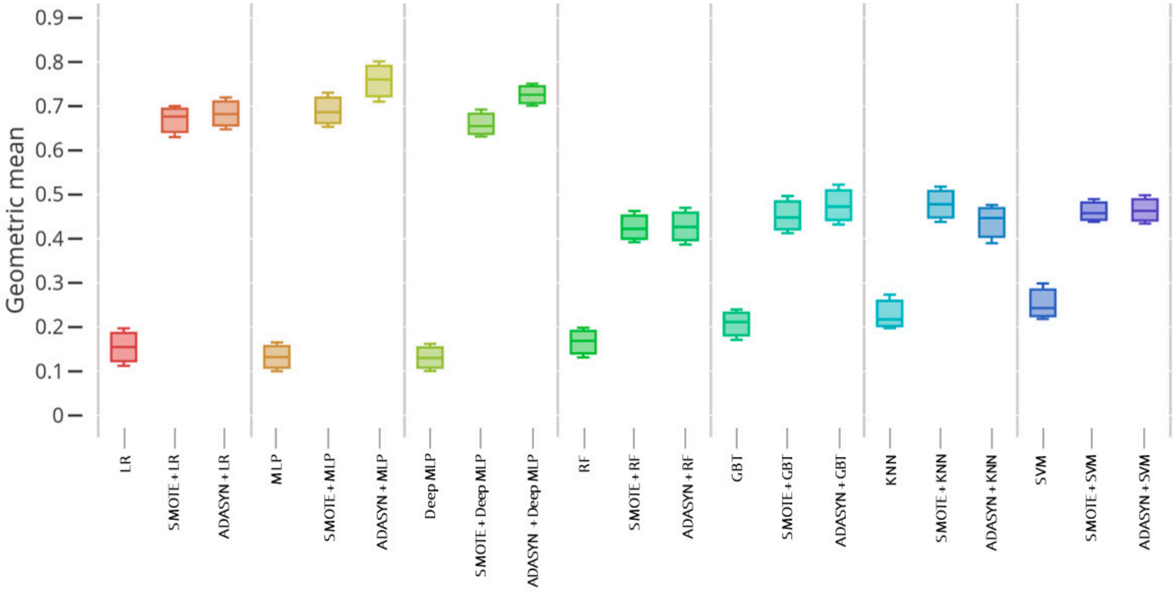

Figure 5. Boxplot of the geometric mean over prediction models among women.

4.4.1. Statistical Test

A non-parametric Friedman test was applied to verify the statistical significance of the differences between the performance results of the various machine learning classifiers. In this study, the prediction models were compared in terms of balanced accuracy score. The Friedman statistic test was calculated as 47.59 with a *p*-value of $1.18 \cdot 10^{-7}$ and rejected the null hypothesis at a 99% significance level. According to the average ranks of the performances, SMOTE with GBT of 2.33 and SMOTE with RF of 3.33 models performed the top rank compared with other models. GBT, LR and DEEP MLP baseline models were determined low ranks by 20.41, 18.91 and 17.08, respectively.

4.4.2. Model Interpretability

Model interpretability is an important task to get a better understanding of the reasoning behind the prediction models. As a comparison result, GBT and RF classifiers determined the best predictive models when combining with the SMOTE among all subjects. Thus, the most important features for SMOTE with GBT and RF models were provided as shown in Figures 6 and 7.

To ensure model interpretability, features were sorted in descending order of their importance scores in model construction. According to the SMOTE with the GBT model, "daily smokers at home", "age of smoking initiation", and "attendance in smoking cessation education" were maintained as the most useful features, with scores of 0.089, 0.088 and 0.079, respectively, to predict the smoking cessation target. On the contrary, "attendance in smoking cessation education", "daily smokers at home" and "household income" features were highly important for constructing the SMOTE with RF model. The highly scored features enhance the rationale decisions in smoking-related health concerns and should be collected in smoking cessation data. Both of the predictive models were constructed by the same features, such as "diabetes" and "hypertension", with low importance scores. Therefore, this analysis is expected to improve the efficiency and effectiveness of healthcare decision support system in real-world adoption.

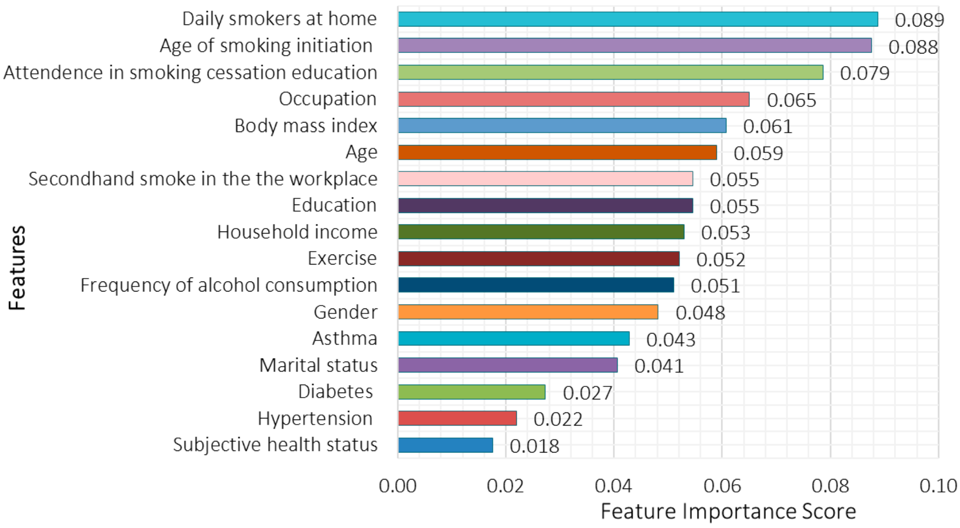

Figure 6. Feature Importance of Synthetic Minority Over-Sampling Technique (SMOTE) with the Gradient Boosting Trees (GBT) model.

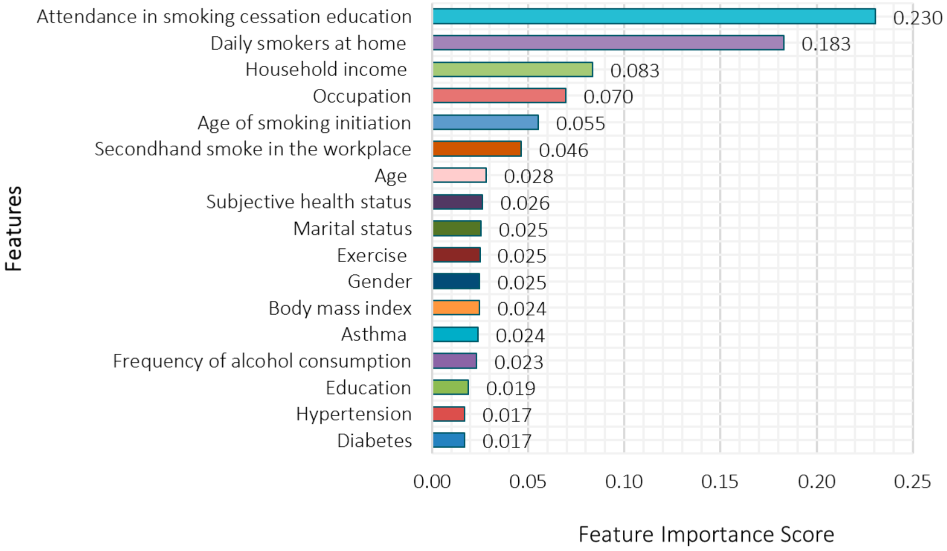

Figure 7. Feature Importance of Synthetic Minority Over-Sampling Technique (SMOTE) with the Random Forest (RF) model.

5. Discussion

Smoking tobacco leads to the occurrence of cancers and premature death. Therefore, secondhand smoke is a serious health hazard that harms for children and nonsmokers. Smoking cessation intervention notably contributes to avoiding smoking prevalence. Previous studies [21–28] have tended to focus on the statistical methods rather than machine learning methods in smoking cessation because of its simplicity. In recent years, few experts have contended that machine learning techniques have been successfully employed to build a prediction of the success of smoking cessation. However, model interpretability tends to be disregarded in those existing studies [29–31] in terms of their complex black-box system. Accordingly, common decision-making responses suffer from the class imbalance problem due to overwhelming skewed distribution. It is well known that higher prediction performance and interpretability are two dominant criteria of the best predictive model.

In this paper, we performed SMOTE and ADASYN techniques for the class imbalance problem in smoking cessation. Our proposed experimental design is composed three components: In the first component, we selected a subset of 17 features with 3692 subjects that were done based on the data preprocessing and feature selection approach that covers lasso and multicollinearity analysis. Selected features were similar to studies [22,31,53] that were utilized in the KNHANES dataset. Aside from accurate models, selecting the representative features is an essential part of the medical domain to understand the significant risk factors. The second component towards solving the class imbalance problem, SMOTE and ADASYN techniques were utilized. In the third component, synthetic oversampling techniques with various machine learning methods were used to construct the prediction models among all subjects, men and women.

According to the comparative results are presented in Tables 3–5, single models were inferior to the synthetic oversampling-based models across all subjects and each gender. Experimental results revealed that the type I error was hard to predict correctly in imbalanced data; this error occurred when successful smoking quitters are misclassified as failed after smoking cessation intervention. Correspondingly, the type II error appeared when current smokers are misclassified as successful quitters. The geometric mean became one of the proper metrics for ascertaining between critical type I and type II errors; thus, we illustrated boxplots of the prediction models using the geometric mean metric, as illustrated in Figures 3–5. As for that boxplot visualization, aspects showed that SMOTE and ADASYN-based classifiers enabled promising performances in each sample.

In addition, model interpretability is another dominant criterion in prediction models. In order to ensure model interpretability, important features were determined based on the best predictive SMOTE-based GBT and RF models. Similar to the study of Blok et al., [28], the "smokers at home" feature was highly important for construct decision support systems in our models. SMOTE with GBT and RF models were provided with similar significant features, such as "attendance in smoking cessation education", "occupation" and "age", the findings of which were similar with studies [22,53]. Aside from accurate model, interpretability refers to obtaining valuable information for the health care experts to make decisions as well as public healthcare concerns.

6. Conclusions

The class imbalance problem is still an open research issue in the machine learning community. In this study, we have presented a comparative study based on the various machine learning methods to predict the success of smoking cessation among the Korean population. In particular, we have utilized the SMOTE and ADASYN techniques for the class imbalance problem. Our experimental analysis was achieved by three components, such as feature selection, data sampling and construct the prediction models among all subjects, men and women. The most important findings revealed that baseline models tend to ignore the minority class, which is leading to higher error rates in imbalanced data. However, comparison results showed the superiority of the synthetic oversampling-techniques-based classifiers in terms of prediction performance. In this study, SMOTE with GBT, RF, MLP provide significantly better handling with regards to the class imbalance problem; moreover, they determined the best prediction across each sample. Furthermore, SMOTE with GBT and RF models provided feature importance scores in order to evaluate the insight of the reasoning behind the models. Overall, we have suggested a comprehensive and interpretative decision support system in public health concerns of the real world.

We have already known that the limitation of this paper is in the fact that collected data provided only former or current smokers after attending smoking cessation intervention. It can be noted that smoking cessation is a slow process because smoking is a highly refractory addiction challenge for smokers. Thereupon, smoking relapse among smokers can be altered in our prediction results. Hence, smokers might be exposed by representative factors that might not have been considered in the dataset of this study. We planned to extend our research by addressing the problem of the interpretability of deep learning models, which supports expert knowledge.

Author Contributions: K.D. and J.S.L. contributed equally to this work. K.D. and J.S.L. conceived to perform the experimental results and wrote the paper. K.D., J.S.L., V.H.P. and K.H.R. discussed and improved the contents of the manuscript. V.H.P. and K.H.R. provided critical insight and discussion. K.H.R. supervised this work. All authors approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: 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.2017R1A2B4010826), (**No.** 2020R1A2B5B02001717), and (No.2019K2A9A2A06020672).

Acknowledgments: The authors would like to thank reviewers for their suggestions to improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ADASYN	Adaptive Synthetic
GBT	Gradient Boosting Trees
KCDC	Korea Centers for Disease Control and Prevention
KNHANES	Korea National Health and Nutrition Examination Survey
KNN	K-Nearest Neighbors
LR	Logistic Regression
MLP	Multilayer Perceptron
RF	Random Forest
SMOTE	Synthetic Minority Over-Sampling Technique
SVM	Support Vector Machine
VIF	Variance Inflation Factor
WHO	World Health Organization

References

1. World Health Organization. *WHO Report on the Global Tobacco Epidemic, 2017: Monitoring Tobacco Use and Prevention Policies*; WHO: Geneva, Switzerland, 2017.
2. WHO Tobacco Free Initiative. *The Role of Health Professionals in Tobacco Control*; World Health Organization: Geneva, Switzerland, 2005.
3. Campion, J.; Checinski, K.; Nurse, J.; McNeill, A. Smoking by people with mental illness and benefits of smoke-free mental health services. *Adv. Psychiatr. Treat.* **2008**, *14*, 217–228. [CrossRef]
4. Song, Y.M.; Sung, J.; Cho, H.J. Reduction and cessation of cigarette smoking and risk of cancer: A cohort study of Korean men. *J. Clin. Oncol.* **2008**, *26*, 5101–5106. [CrossRef] [PubMed]
5. Li, X.H.; An, F.R.; Ungvari, G.S.; Ng, C.H.; Chiu, H.F.; Wu, P.P.; Xiang, Y.T. Prevalence of smoking in patients with bipolar disorder, major depressive disorder and schizophrenia and their relationships with quality of life. *Sci. Rep.* **2017**, *7*, 8430. [CrossRef] [PubMed]
6. Milcarz, M.; Polanska, K.; Bak-Romaniszyn, L.; Kaleta, D. Tobacco Health Risk Awareness among Socially Disadvantaged People—A Crucial Tool for Smoking Cessation. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2244. [CrossRef] [PubMed]
7. Yang, H.K.; Shin, D.W.; Park, J.H.; Kim, S.Y.; Eom, C.S.; Kam, S.; Seo, H.G. The association between perceived social support and continued smoking in cancer survivors. *Jpn. J. Clin. Oncol.* **2012**, *43*, 45–54. [CrossRef]
8. Rigotti, N.A. Strategies to help a smoker who is struggling to quit. *JAMA* **2012**, *308*, 1573–1580. [CrossRef]
9. Hyndman, K.; Thomas, R.E.; Schira, H.R.; Bradley, J.; Chachula, K.; Patterson, S.K.; Compton, S.M. The Effectiveness of Tobacco Dependence Education in Health Professional Students' Practice: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4158. [CrossRef]
10. Kim, H.; Ishag, M.; Piao, M.; Kwon, T.; Ryu, K.H. A data mining technique for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries. *Symmetry* **2016**, *8*, 47. [CrossRef]
11. Lee, H.C.; Yoon, H.K.; Nam, K.; Cho, Y.; Kim, T.; Kim, W.; Bahk, J.H. Derivation and validation of machine learning techniques to predict acute kidney injury after cardiac surgery. *J. Clin. Med.* **2018**, *7*, 322. [CrossRef]
12. Heo, B.M.; Ryu, K.H. Prediction of Prehypertenison and Hypertension Based on Anthropometry, Blood Parameters, and Spirometry. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2571. [CrossRef]

13. Yang, E.; Park, H.; Choi, Y.; Kim, J.; Munkhdalai, L.; Musa, I.; Ryu, K.H. A simulation-based study on the comparison of statistical and time series forecasting methods for early detection of infectious disease outbreaks. *Int. J. Environ. Res. Public Health* **2018**, *15*, 966. [CrossRef] [PubMed]
14. Zhu, M.; Xia, J.; Jin, X.; Yan, M.; Cai, G.; Yan, J.; Ning, G. Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access* **2018**, *6*, 4641–4652. [CrossRef]
15. Dal Pozzolo, A.; Caelen, O.; Le Borgne, Y.A.; Waterschoot, S.; Bontempi, G. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Syst. Appl.* **2014**, *41*, 4915–4928. [CrossRef]
16. Le, T.; Lee, M.; Park, J.; Baik, S. Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset. *Symmetry* **2018**, *10*, 79. [CrossRef]
17. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
18. Douzas, G.; Bacao, F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Syst. Appl.* **2018**, *91*, 464–471. [CrossRef]
19. Dobbins, C.; Rawassizadeh, R.; Momeni, E. Detecting physical activity within lifelogs towards preventing obesity and aiding ambient assisted living. *Neurocomputing* **2017**, *230*, 110–132. [CrossRef]
20. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887.
21. Monso, E.; Campbell, J.; Tønnesen, P.; Gustavsson, G.; Morera, J. Sociodemographic predictors of success in smoking intervention. *Tob. Control* **2001**, *10*, 165–169. [CrossRef]
22. Kim, Y.J. Predictors for successful smoking cessation in Korean adults. *Asian Nurs. Res.* **2014**, *8*, 1–7. [CrossRef] [PubMed]
23. Charafeddine, R.; Demarest, S.; Cleemput, I.; Van Oyen, H.; Devleesschauwer, B. Gender and educational differences in the association between smoking and health-related quality of life in Belgium. *Prev. Med.* **2017**, *105*, 280–286. [CrossRef] [PubMed]
24. Lee, S.; Yun, J.E.; Lee, J.K.; Kim, I.S.; Jee, S.H. The Korean prediction model for adolescents' future smoking intentions. *J. Prev. Med. Public Health* **2010**, *43*, 283–291. [CrossRef] [PubMed]
25. Kim, S.H.; Lee, J.A.; Kim, K.U.; Cho, H.J. Results of an inpatient smoking cessation program: 3-month cessation rate and predictors of success. *Korean J. Fam. Med.* **2015**, *36*, 50. [CrossRef] [PubMed]
26. Foulds, J.; Gandhi, K.K.; Steinberg, M.B.; Richardson, D.L.; Williams, J.M.; Burke, M.V.; Rhoads, G.G. Factors associated with quitting smoking at a tobacco dependence treatment clinic. *Am. J. Health Behav.* **2006**, *30*, 400–412. [CrossRef] [PubMed]
27. Smit, E.S.; Hoving, C.; Schelleman-Offermans, K.; West, R.; de Vries, H. Predictors of successful and unsuccessful quit attempts among smokers motivated to quit. *Addict. Behav.* **2014**, *39*, 1318–1324. [CrossRef]
28. Blok, D.J.; de Vlas, S.J.; van Empelen, P.; van Lenthe, F.J. The role of smoking in social networks on smoking cessation and relapse among adults: A longitudinal study. *Prev. Med.* **2017**, *99*, 105–110. [CrossRef]
29. Coughlin, L.N.; Tegge, A.N.; Sheffer, C.E.; Bickel, W.K. A machine-learning technique to predicting smoking cessation treatment outcomes. *Nicotine Tob. Res.* **2018**. [CrossRef]
30. Poynton, M.R.; McDaniel, A.M. Classification of smoking cessation status with a backpropagation neural network. *J. Biomed. Inform.* **2006**, *39*, 680–686. [CrossRef]
31. Davagdorj, K.; Yu, S.H.; Kim, S.Y.; Huy, P.V.; Park, J.H.; Ryu, K.H. Prediction of 6 Months Smoking Cessation Program among Women in Korea. *Int. J. Mach. Learn. Comput.* **2019**, *9*, 83–90. [CrossRef]
32. Meier, L.; Van De Geer, S.; Bühlmann, P. The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2008**, *70*, 53–71. [CrossRef]
33. Salmerón Gómez, R.; García Pérez, J.; López Martín, M.D.M.; García, C.G. Collinearity diagnostic applied in ridge estimation through the variance inflation factor. *J. Appl. Stat.* **2016**, *43*, 1831–1849. [CrossRef]
34. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
35. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling technique for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
36. Bagley, S.C.; White, H.; Golomb, B.A. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *J. Clin. Epidemiol.* **2001**, *54*, 979–985. [CrossRef]

37. Lisboa, P.J. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Netw.* **2002**, *15*, 11–39. [CrossRef]
38. Mazurowski, M.A.; Habas, P.A.; Zurada, J.M.; Lo, J.Y.; Baker, J.A.; Tourassi, G.D. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Netw.* **2008**, *21*, 427–436. [CrossRef]
39. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
40. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
41. Tan, P.N. *Introduction to Data Mining, Pearson Education India*; Indian Nursing Council: New Delhi, India, 2018.
42. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. TIST* **2011**, *2*, 27. [CrossRef]
43. Zeng, X.; Martinez, T.R. Distribution-balanced stratified cross-validation for accuracy estimation. *J. Exp. Theor. Artif. Intell.* **2000**, *12*, 1–12. [CrossRef]
44. Benavoli, A.; Corani, G.; Mangili, F. Should we really use post-hoc tests based on mean-ranks? *J. Mach. Learn. Res.* **2016**, *17*, 152–161.
45. Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Proceedings of the European Conference on Information Retrieval, Santiago de Compostela, Spain, 21–23 March 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 345–359.
46. Altman, D.G.; Bland, J.M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ Br. Med. J.* **1994**, *308*, 1552. [CrossRef]
47. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 559–563.
48. McKinney, W. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; Volume 445, pp. 51–56.
49. Oliphant, T.E. *A guide to NumPy*; Trelgol Publishing: Provo, UT, USA, 2006; Volume 1, p. 85.
50. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90. [CrossRef]
51. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; van der Walt, S.J. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
52. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Vanderplas, J. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, 2825–2830.
53. Davagdorj, K.; Lee, J.S.; Park, K.H.; Ryu, K.H. A machine-learning approach for predicting success in smoking cessation intervention. In Proceedings of the 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 23–25 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.