Prediction of 6 Months Smoking Cessation Program among Women in Korea

Khishigsuren Davagdorj, Seon Hwa Yu, So Young Kim, Pham Van Huy, Jong Hyock Park, and Keun Ho Ryu

***Abstract*—Cigarette smoking is the leading cause of preventable death in a general population and it seems a significant topic in health research. The primary aim of this study determines the significant risk factors and investigates the prediction of 6 months smoking cessation program among women in Korea. In this regard, we examined real-world dataset about a smoking cessation program among the only women from Chungbuk Tobacco Control Center of Chungbuk National University College of Medicine in South Korea which collected from 2015 to 2017. Accordingly, we carried out to compare four machine learning techniques: Logistic regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Naïve Bayes (NB) in order to predict response for successful or unsuccessful smoking quitters. Totally we analyzed 60 set of features that may affect the association between smoking cessation such as socio-demographic characteristics, smoking status for the age of starting, duration and others by employing a filter-based feature selection method. Respectively, we identified significant 8 factors which associated with smoking cessation. The experimental results demonstrate that NB performs better than other classifiers. Moreover, the performance of prediction models as measured by Accuracy, Precision, Recall, F-measure and ROC area. This finding has gone some way towards enhancing our better understanding of the significant factors contributing to smoking cessation program implementation and accompanying to concern public health.**

***Index Terms*—Smoking cessation, women, feature selection, logistic regression, support vector machine, random forest, Naïve Bayes.**

Manuscript received August 25, 2018; revised November 3, 2018. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-2013-1-00881) supervised by the IITP (Institute for Information & communication Technology Promotion) and 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) and conducted with the support of the National Health Promotion Fund. This paper has been deliberated by IRB.

Khishigsuren Davagdorj is with the Database/Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, South Korea (e-mail: suri@dblab.chungbuk.ac.kr).

Seon Hwa Yuand So Young Kim are with the Chungbuk Tobacco Control Center, Chungbuk National University, South Korea (email: tocjstk1256@gmail.com, sykim@gmail.com)

Pham Van Huy is with the Faculty of Information Technology of Ton Duc Thang University, Vietnam (e-mail: phamvanhuy@tdt.edu.vn)

Jong Hyock Park is with the Chungbuk National University College of Medicine, South Korea, and Chungbuk Tobacco Control Center, Chungbuk National University, South Korea (email: jonghyock@gmail.com)

Keun Ho Ryu is with the Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam and Department of Computer Science, College of Electrical and Computer Engineering, Chungbuk National University, South Korea (corresponding author: Keun Ho Ryu; e-mail: khryu@tdtu.edu.vnr, khryu@.chungbuk.ac.kr).

I. Introduction

Tobacco use is the widely documented preventable risk factor for premature death as it kills about more than 5 million people throughout worldwide in every year. 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 reaching 8 million per year. Moreover, largely the growth of over 80% of tobacco smokers live in low and middle-income countries [1], [2].

Most smokers want to quit smoking, furthermore as known as majority make multiple quit attempts during their lifetime but many people eventually failed in smoking cessation [3]. The reason for these spread of critical evidence, increase the awareness about the impact of smoking dangers on health and aware the antismoking legislation in order to prevention policies for offering quit smoking in social. Many countries have been realizing to decrease tobacco consumption through monitoring and implementing smoke-free ways for encouraging smokers to quit effectively. Especially, government and health care providers initiate to implement more accessible resources to help smokers to quit.

In point of fact, Tobacco Control Center was established in 18 cities of the Republic of Korea from 2015. An important component of smoking cessation program is understanding the factors and predicting success for quitting which is an effective way for public health benefit.

According to a report from the World Health Organization, women have traditionally not used tobacco permanently as well women smoke at about one fourth the rate of men. Even though, compare non-smoker women with a smoking dependent women who has the greater risk of reproductive health problems, many forms of gynecologic and other types of cancer, coronary and vascular disease, chronic obstructive lung disease, and osteoporosis [4].

A recent literatures [5]-[7] in this area examining factors associated with smoking cessation based on sampled population, for example, participants of smoking cessation intervention defined period or certain generalizations of group objects.

S. Kim [5] study evaluated smoking prevalence for Korean adults by gender, age group and the association between smoking and socio-demographic factors using the Korea National Health and Nutrition Examination Survey (KNHNES) 2008-2010 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

doi: 10.18178/ijmlc.2019.9.1.769

and confidence intervals in 95% for finding association and comparison of variables.

R. Charafeddine *et al*. [6] 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 significantly lower health-related quality of life scores compared with never smokers, but only among females with a low and intermediate educational level.

I. Khati *et al*. [7] compared individuals who successfully quit smoking from those who relapsed on socio-demographic, psychological and health factors based on data coming from telephone interviews conducted in 2011 with participants of the TEMPO (Trajectoires Epid é Miologiques en Population). They conducted the regression analyses and multivariate analyses within a stepwise descending method. Their result shows that 43% of participants were current smokers who never quit for the extended period and, 33% former smokers and 24% current smokers who relapsed after extended cessation. Therefore, they concluded about work and family circumstances, co-occurring substance use and psychological difficulties might affect smoking cessation in young adults.

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 estimating statistical significant interactions among dependent variable and one or more independent variables.

Nowadays classification technique plays an essential role in drive the decision rules effectively. 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. Data classification process consist of two-steps such as building the model and using the classification model for classification. While step of building the model, the classification model is constructed by a predetermined training set, subsequently applied it to the test set which consists of records with unknown class labels.

Varies application motivated by the success of the classification techniques, especially in the medical domain [8]-[10] utilized widely. Therefore, an objective of these designed built to compute the classifiers evaluation, in the result, explore the best models for supporting their decision.

The organization of the experimental steps are as follows: Our proposed framework has three main components: First, we analyze data preprocessing and determine significant features. Second, apply to compare the results of Logistic Regression (LR), Support vector machine (SVM), Random Forest (RF) and Na ïve Bayes (NB). Final step is performance evaluation, we will propose the best prediction model in smoking cessation result only women after 6 months program.

The remainder of this paper is logically structured as follows: Section II describes a dataset, feature selection and classification methods we used. Framework and experimental result demonstrates in Section III. Finally, conclusion and future work are presents in Section IV.

II. Materials and Methods

A. Data Interpretation

This study examined real-world data from Chungbuk Tobacco Control Center of Chungbuk National University College of Medicine in South Korea which collected from 2015 to 2017. The current study was approved by the Institutional Review Board (IRB) of the Chungbuk National University (IRB approval No.CBNU-201801-SBETC-591-01).

In this study, we evaluated only about smoking relapse among women through participation of 6 months smoking cessation program. Prospective sampled raw data contains 60 features and 407 women who cigarette smoking.

B. Feature Selection and Creation

Feature selection [11] is an essential preprocessing step in data mining for selecting a subset of relevant features and improving performance for classifiers from the original dataset. Although, feature selection method eliminates redundant and irrelevant features order to distinguish features with which it is higher correlated. Feature selection method can be categorized filter, wrapper and embedded approach illustrated in Fig. 1.

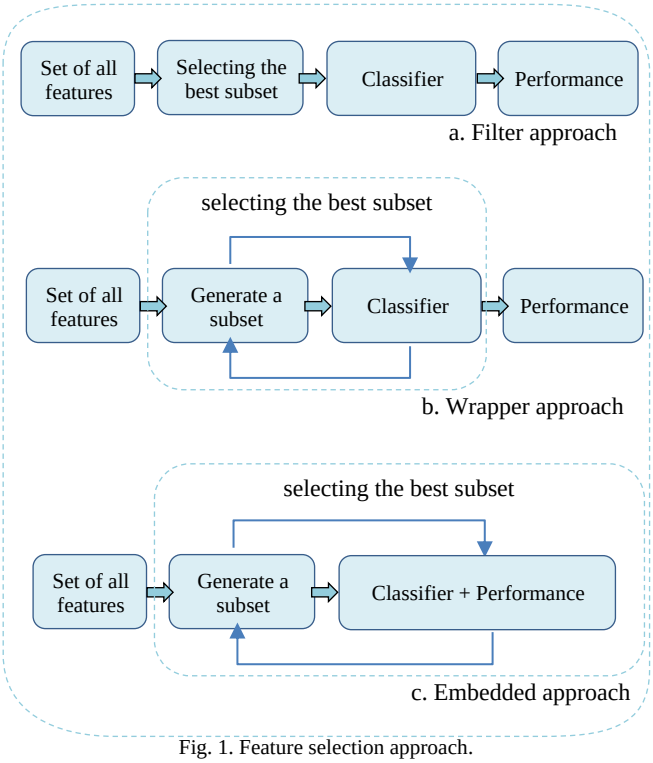

Fig. 1. Feature selection approach.

Filter approach is generally do not require any classification algorithm. Moreover, the filter approach is faster in computation time and scalable to high dimensional data. For this reason, we applied the filter feature selection approach in this paper.

Wrapper approach is not same with filter approach because it detects the possible interactions between feature subsets. The main disadvantage of the wrapper approach is high time computation when large data and has a risk of overfitting. Embedded approach is to combine the filter and wrapper approaches and can cover high dimensional data as well. But wrapper and embedded approaches have the same drawback which is classifier dependent selection. [12], [13].

Feature creation has methodologies for extracting a new set of attributes, mapping the data to a new space and constructing to provide necessary information and in some cases leading to better domain understanding.

C. Logistic Regression (LR)

LR is a statistical method for analyzing a dataset where the dependent variable is categorical. 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 logit function determined as the natural logarithm (ln) of the "odds" of the target variable, used to "S" shaped curve bounded to be between (0 and 1) to a variable that ranges over $(-\infty, +\infty)$ [14], [15]. The LR model is:

$$logit(p) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k \quad (1)$$

where p is the probability of presence of the characteristic for interest and logged odds is defined by:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probabiltiy\ of absence\ of\ characteristic} \quad (2)$$

The logistic formulas are existed in term of the probability that $Y = 1$ is yes and $Y = 0$ no means $1 - p$.

$$\ln\left(\frac{p}{1-p}\right) = \beta \cdot X_i \quad (3)$$

where $\beta \cdot X_i$ is familiar equation with linear regression line and it suspects form the distribution $P(Y|A)$ and parametric model is:

$$P(Y = yes|A) = \frac{\exp(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)}{1+\exp(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)} \quad (4)$$

and therefore,

$$P(Y = no|A) = \frac{1}{1+\exp(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)} \quad (5)$$

where β_i - is the coefficient of the predictor variable and slope can be interpreted as the change of Y, from unit change in X [10]. The LR model can be expressed as follows:

$$ln\left(\frac{F(x)}{1-F(x)}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \quad (6)$$

where $(F(x))$ – probability of prediction, β_0 – constant coefficient , β_i – coefficient corresponding to the feature x_i

D. Support Vector Machine (SVM)

Support Vector Machine (SVM) [16], [17] is an Artificial Intelligence-based technique which can be classification and regression problems.

SVM to find the decision boundary with maximal margin where the distance between two groups of data points. Here, SVM search an optimal separating hyperplane which divides two classes correctly. Based on the features of support vectors, which suitable belongs to classes as successful smokers or unsuccessful smokers can be predicted. The main objective of SVM is to map the original training set to high-dimensional feature space as shown in Fig. 2.

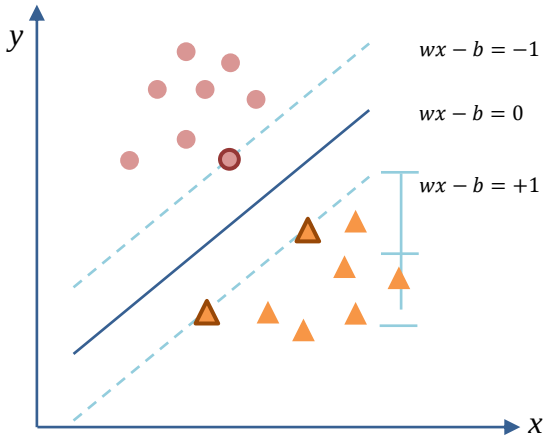

Fig. 2. Illustration of support vector machine.

SVM function is formulated as follows:

$$f(x) = w^T \varphi(x) + b \quad (7)$$

where w is a vector weight coefficient, $\varphi(x)$ represents a vector in the corresponding high - dimensional space comprising nonlinear attributes and b is bias constant. w and b are estimated by minimizing the following optimization problem:

$$minimize\ \frac{1}{2}\|w\|^2 \quad (8)$$

$$subject\ to\ \begin{cases} y_i - (\langle w, \varphi(x_i)\rangle + b) \le \varepsilon \\ (\langle w, \varphi(x_i)\rangle + b) - y_i \le \varepsilon \end{cases} \quad (9)$$

where ε is a free parameter that serves as a threshold: all predictions have to be within an ε range of the true predictions. Slack variables are usually added into the above to allow for errors and to allow approximation in the case the above problem is infeasible.

E. Naïve Bayes Classifier (NB)

Naive Bayesian [18], [19] is statistical classifier assume that the attributes are conditionally independent, given the particular class label (successful quitter or unsuccessful quitters for smoking). This classifier named by class-conditional independence and based on Bayes's Theorem. NB classifiers examine the notion of conditional probability

Let X, Y and Z denote three sets of random variables. The variables in X are expressed to be conditionally independent of Y, given Z, if the following condition holds:

$$P(X|Y, Z) = P(X|Z) \quad (10)$$

The conditional independence between X and Y can also be written into a form that as:

This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample. Mathematically this means formula can be written as:

$$P(X|Y, Z) = \frac{P(X,Y,Z)}{P(Z)} = P(X|Z) \times P(Y|Z) \quad (11)$$

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|c_i) \quad (12)$$

where the probability $P(x_1|C_i), P(x_2|C_i), \ldots, P(x_n|C_i)$ be estimated from the training set. x_k refers to the value of attribute.

F. Random Forest (RF)

Random Forest (RF) [20] is a class ensemble tree-based method which bagging to generate subsets of the entire training set to build multiple individual decision trees as shown in Fig. 3.

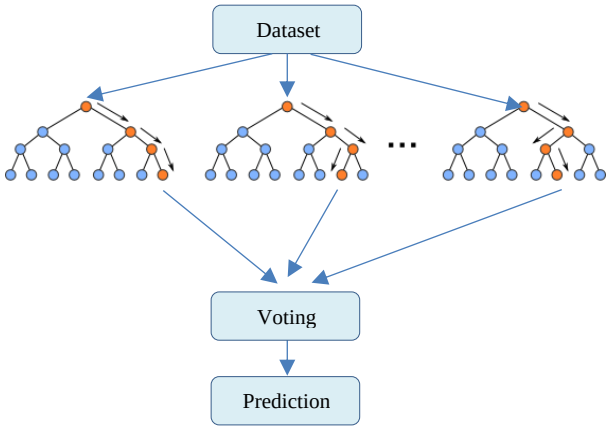

Fig. 3. Random forest.

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 RF over single decision tree classifier is reducing over-fitting of training data and get more accurate. The reason of it, we used RF ensemble method for predicting success or unsuccessful reason for smokers as well.

G. Evaluation Metrics

Evaluation of the performance [21], [22] of each classification model is evaluated using measures such as accuracy, precision, recall, F-measure and receiver operating characteristic (ROC) area.

These classification measures are determined using four value, namely true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Correct or incorrect classified instances predicted by the model and these counts are known as a confusion matrix for a binary classification problem which illustrated by Table I. Based on the entries in the confusion matrix, total number of (TP + TN) can interpret correct predictions and incorrect predictions built by the model is (FN + FP) respectively.

TABLE I: CONFUSION MATRIX FOR A TWO CLASS PROBLEM

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

As a consideration of this provided information of confusion matrix, performance metric such as accuracy, precision, recall, F-measure and ROC area which are computed by Eq.13-16.

Accuracy is defined as the overall success rate of the classifier and is equal to the sum of TP and TN divided by total number of entries.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (13)$$

Precision measures the fraction of true or correctly classified point pairs compared to all the point pairs in the same class.

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

Recall measures the fraction of correctly classified points pairs compared to all the point pairs in the same class.

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

F-measure is harmonic mean of precision and recall.

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (16)$$

ROC area is a probablity curve and plotted with TP rate against the FP rate where TP rate is on y-axis and FP rate is on x-axis.

III. EXPERIMENT AND RESULT

Our workflow of the experiment was illustrated in Fig. 4.

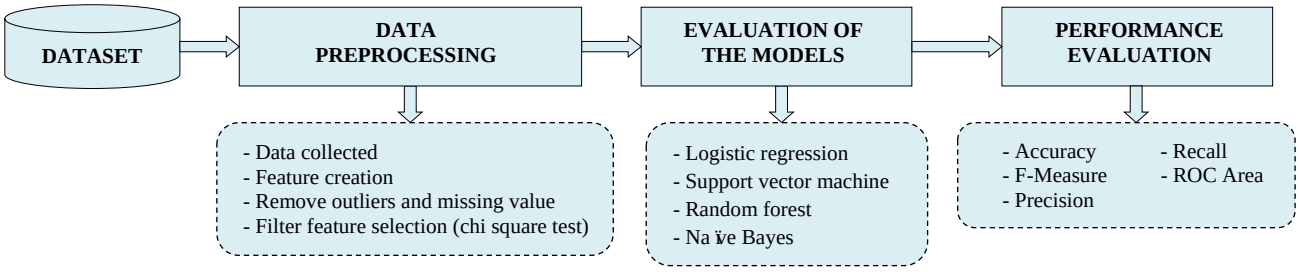

Fig. 4. Workflow of the experiment.

Firstly, in the pre-processing step, we discretized for continues data based on the quartile-based method and selected the significant features in smoking cessation using chi-square test. If it equal or less than 0.05, applied into the second step for applying proposed comparing algorithms.

Secondly, we applied four classification algorithms: LR, SVM, RF and NB with 10 cross-validation method.

Then, we evaluated to compare the performances by

accuracy, precision, recall, F-measure and ROC area, and these performance measures are defined as a confusion matrix which described a difference between the actual and predicted values of variables.

A. Preprocessing

In our experiment, data-preprocessing is generated general four steps and its process summarized in Fig. 5.

The first step, we collected the sampled raw data which contains totally 60 features and the total number of study subjects were 407 women who participated in the smoking cessation program.

The second step, we investigated feature creation and quartile-based discretization method that depends on distribution for continues data through discussing with specialists of Chungbuk Tobacco Control Center. The result of this step, we generated 18 features which can be express the implementation of the program and quit smoking initiative.

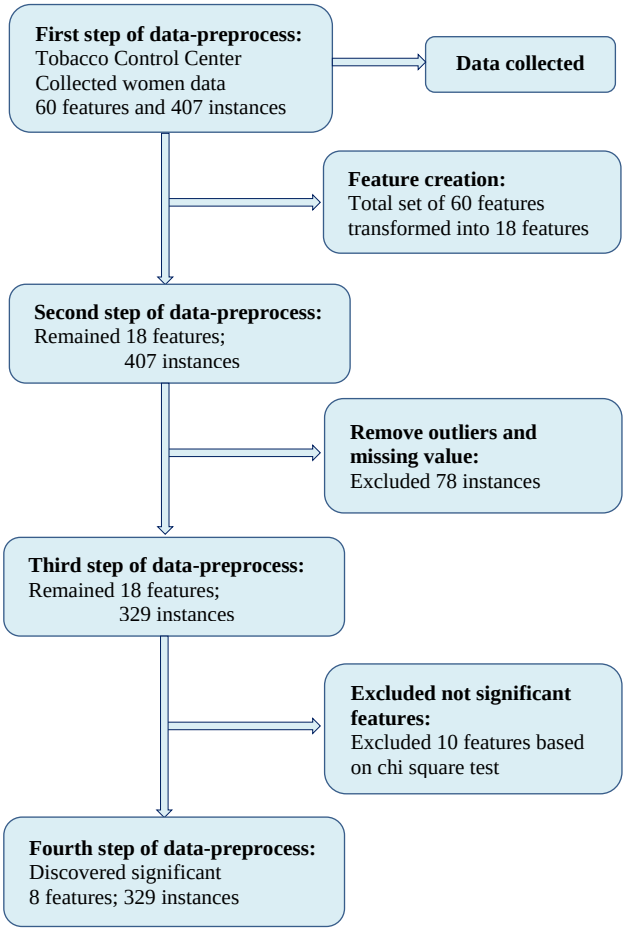

Fig. 5. Data generation process.

The third step, all of the outliers and missing values were removed in order to find good quality of the result.

The fourth step, we analyzed the significance of each attributes with 6 months smoking cessation using a chi-square test for categorized features respectively. Significant filter feature selection based on chi-square test flowchart shown in Fig. 6.

Age, counseling frequency, exhalation carbon monoxide concentration, age at smoking initiation, duration of smoking by year and number of cigarettes smoked per day features were calculated mean and standard deviation is shown in Table II. We determined p value equal or less than 0.05 to indicate strong evidence which ignores the null hypothesis as well as known considers age, registration motive, medical guarantee, medical condition, body mass index, counseling frequency, exhalation carbon monoxide concentration and age at smoking initiation which were highly correlated with smoking quitters of success rate.

Finally, after all of the steps of data pre-processing, preprocessed data has 329 instances and 8 features to forward in next classification analyze.

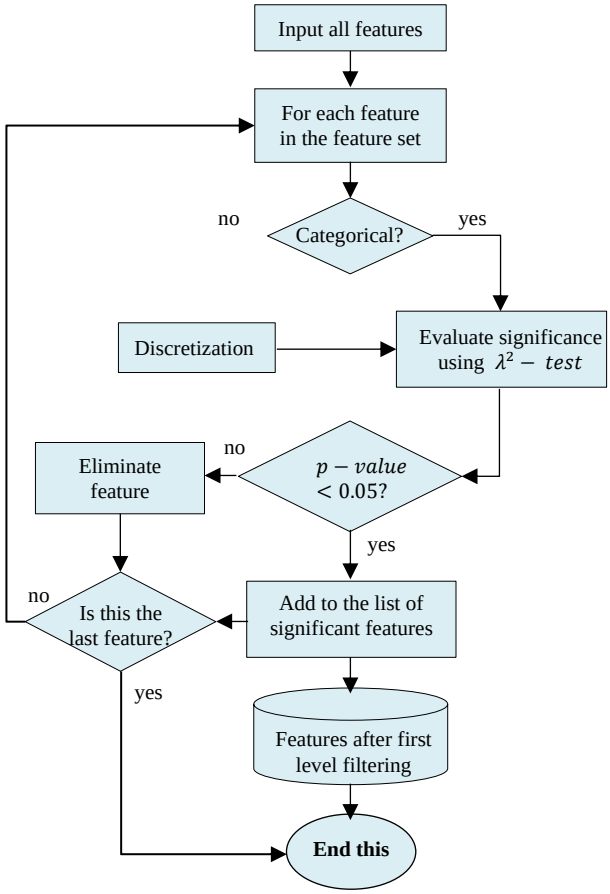

Fig. 6. Significant feature selection based on chi-square.

TABLE II: BIVARIATE ANALYSIS OF SELECTED CHARACTERISTICS BY SMOKING STATUS AFTER 6 MONTHS

Variables	Successful quitters (%)	Unsuccessful quitters (%)	$\chi^2(\rho)$
Age			
< 39	25 (10.3)	218 (89.7)	9.759
40-54	6 (9.8)	55 (90.2)	(0.021)
55-64	14 (25.0)	42 (75.0)	
>=65	5 (10.6)	42 (89.4)	
Total	50 (12.3)	357 (87.7)	
M±SD	42.64±17.033	39.42±18.527	
Education†			
Up to high school graduate	30 (10.0)	266 (89.9)	3.423 (0.064)
College graduate or higher	15 (17.4)	71 (82.6)	
Total	45 (11.7)	337 (88.2)	
Occupation			
Manager, specialist or office worker	3 (17.6)	14 (82.4)	4.302 (0.331)*
Service or seller	24 (11.3)	189 (88.7)	
Function, device machine assembly worker or farmers	1 (10.0)	9 (90.0)	
Non economically active population (including students)	10 (20.4)	39 (79.6)	

Others	12 (10.2)	106 (89.8)	
Total	50 (12.3)	357 (87.7)	
Registration motive			
TV or advertisement	14 (29.2)	34 (70.8)	16.299
Visual media (banner, poster, promotional books, activity or event)	4 (4.9)	78 (95.1)	(0.004)*
Internet or notice of public health center	3 (18.8)	13 (81.2)	
Invitation of neighbor (neighbor, smoking cessation counseling call or medical team)	5 (9.1)	50 (90.9)	
Multiple response	1 (12.5)	7 (87.5)	

Fig. 5. Data generation process

Total	50 (12.3)	357 (87.87)	
Other smokers in environment			
Yes	38 (13.3)	248 (86.7)	0.896
No	12 (9.9)	109 (90.1)	(0.344)
Total	50 (12.3)	357 (87.7)	
Medical guarantee			
Medical benefits	1 (7.7)	12 (92.3)	8.645
Health insurance	40 (16.1)	209 (83.9)	(0.011)*
Others	9 (6.2)	136 (93.8)	
Total	50 (12.3)	357 (87.7)	
Medical conditions			
Hypertension	8 (44.4)	10 (55.6)	17.629
Diabetes	0 (0.0)	5 (100)	(0.002)*
Hyperlipidemia	0 (0.0)	5 (100)	
Others	1 (7.1)	13 (92.9)	
2 or more disease	35 (10.2)	307 (89.8)	
None	6 (26.1)	17 (73.9)	
Total	50 (12.3)	357 (87.7)	
Body mass index†			
Underweight (13-18.5)	2 (5.0)	38 (95.0)	5.869
Average (18.5~25.0)	30 (11.2)	237 (88.8)	(0.050)*
Overweight (26.0-40.0)	11 (21.6)	40 (78.4)	
Total	43 (12.0)	315 (88.0)	
Exercise			
Yes	15 (15.3)	83 (84.7)	1.093
No	35 (11.3)	274 (88.7)	(0.296)
Total	50 (12.3)	357 (87.7)	
Frequency of alcohol consumption in recent 1 year			
Yes	33 (12.4)	234 (87.6)	0.004
No	17 (12.1)	123 (87.9)	(0.950)
Total	50 (12.3)	357 (87.7)	
Counseling frequency			
<=2	0 (0.0)	151 (100)	106.378
3-5	0 (0.0)	114 (100)	(0.000)
>=6	50 (35.2)	92 (64.8)	
Total	50 (12.3)	357 (87.7)	
M±SD	10.02±3.172	4.05±3.031	
Exhalation carbon monoxide concentration (ppm)†			
0-6	33 (19.0)	141 (81.0)	13.387
7-10	9 (11.2)	71 (88.8)	(0.007)*
11-15	5 (7.0)	66 (93.0)	
16-25	2 (3.3)	58 (96.7)	
26-50	1 (6.2)	15 (93.8)	
Total	50 (12.5)	351 (87.5)	
M±SD	6.04±5.883	9.76±7.856	
Age at smoking initiation†			
<19	22 (11.2)	174 (88.8)	8.472
20-29	12 (10.3)	104 (89.7)	(0.037)
30-39	11 (26.2)	31 (73.8)	
>=40	5 (11.2)	174 (88.8)	
Total	50 (12.3)	356 (87.7)	
M±SD	24.60±11.317	24.29±13.215	
Duration of smoking (year)†			
1-19	31 (10.6)	261 (89.4)	3.494
20-29	11 (19.3)	46 (80.7)	(0.174)
>=30	7 (14.0)	43 (86.0)	
Total	49 (12.3)	350 (87.7)	
M±SD	16.72±11.227	14.39±11.259	
No. of cigarettes smoked per day			
0-10	35 (13.8)	219 (86.2)	1.558
11-19	5 (8.5)	54 (91.5)	(0.459)
>=20	10 (10.6)	84 (89.4)	
Total	50 (12.3)	357 (87.7)	
M±SD	10.90±7.571	12.07±8.317	
Provided nicotine supplement or whether†			
No	26 (13.5)	166 (86.5)	0.324
Yes	24 (11.7)	182 (88.3)	(0.569)
Total	50 (12.6)	357 (87.4)	
Provided behavioral therapy or whether (vitamin, menthol etc.)†			
No	11 (9.8)	101 (90.2)	1.066
Yes	39 (13.6)	247 (86.4)	(0.302)
Total	50 (12.6)	348 (87.4)	
Nicotine dependence			
Low (0-3)	30(14.3)	180 (85.7)	1.646
Medium (4-6)	15(10.4)	129(89.6)	(0.439)
High (>=7)	5(9.4)	48(90.6)	
Total	50(12.3)	357(87.7)	

* - Fisher exact test

† - excluded missing value and outlier

M - Mean

SD – Standard Deviation

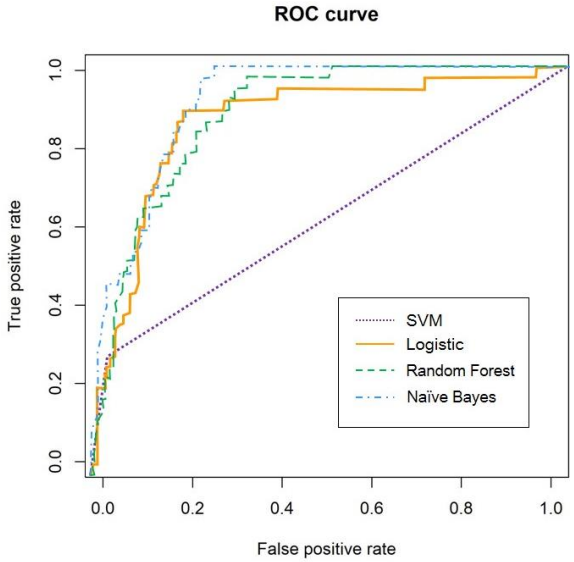

Fig. 7. Comparison of ROC area of classifiers.

B. Experimental Evaluation

In this section, we describe and compute the performance in machine learning algorithms by employing 10 fold cross-validation method. The results are summarized in Table III, which present that the NB classifier based on filter feature selection method achieves encouraging performances across our analyzing dataset. The best run performances are in bold for each measure.

NB classifier model outperforming in evaluation measures such as accuracy (90.2%), Precision (88.9%) Recall (90.3%) and F-measure (89.1%) respectively among the four algorithms have experimented. Especially ROC area which defined by True positive and False positive rate of predicted value in actual value evaluated by 91.1%. On the contrary, compared with among all classifiers RF performed slightly less performance for prediction accuracy (87.2%), Precision (81.4%) Recall (87.2%), F-measure (83.4%) and ROC Area (66.6%) in our experiment.

Indicating some error distributions are remarkable equal for LR and SVM. The second-lowest resulted benchmark model is LR in accuracy (87.5%), Precision (86.5%) Recall (87.5%), F-measure (87.0%) and excluding ROC area (86.5). Support vector machine examined worst ROC Area

performance (62.8%) compare with proposed algorithms. Comparison of ROC area measurements as illustrated in **Fig. 7.**

TABLE III: COMPARISON RESULTS OF PREDICTION MODEL FOR SMOKING CESSATION PROGRAM AMONG WOMEN IN EACH ALGORITHMS

	Logistic Regression	Support Vector Machine	Random Forest	Naïve Bayes
Accuracy (%)	87.5	88.7	87.2	**90.2**
Precision (%)	86.5	86.7	81.4	**88.9**
Recall (%)	87.5	88.8	87.2	**90.3**
F-Measure (%)	87.0	87.3	83.4	**89.1**
ROC Area (%)	86.5	62.8	66.6	**91.1**

In sum, in terms of the given imbalanced dataset, NB dominated to perform the best model. Thus, the SVM classifier can predict it adequately in accuracy, precision, recall and F-Measure, whereas evaluating the ROC area measure by useless in our dataset.

These experimental results lead us to new directions for the prediction model for tobacco-dependent women who participated in the smoking cessation program through 6 months. Even 88.7% of smokers cannot quit unsuccessfully, our discovered significant features and model would provide to understand about this area and implement this required program more effectively.

IV. CONCLUSION AND FURTHER WORK

In this study, we collected smoking cessation program among women who controlled by Chungbuk Tobacco Control Center about 6 months. In the preprocessing step, we discovered significant features for interrupting in smoking relapse for women through analyzed by statistic hypothesis chi-square test from discretized features. We purposed also a better understanding of the factors contributing to relapse smoking could be a contribution for implementing this kind program and protect the health of the public.

Despite the fact that, we adopted machine learning algorithms such as LR, SVM, RF and NB based in filter feature selection method for designing prediction model for smoking cessation program. One of the more significant finding to emerge from this study is that represents that, NB algorithm has the best performances among all classifiers while analyzing the imbalanced dataset. This finding has gone some way towards enhancing our understanding of prediction in this area. Even, 88.7% of our analyzing objectives failed smoking cessation program while participating 6 months smoking cessation program along with several related risk factors dependence for counseling frequency and age respectively. Moreover, objectives who has a disease such as hypertension, diabetes and hyperlipidemia were also less likely to quit unsuccessfully.

Accordingly, our finding suggest a cessation program that considering these finding in setting up based on patients condition.

The generalizability of these results is subject to certain limitations. For instance, we didn't analyze broadly to compare with other objectives and comparing more method and algorithms of machine learning yet. Further experimental investigations are remained to estimate the limited works of this study and finding associative rules, especially disease.

ACKNOWLEDGMENT

The authors would like to thank Jung Rak Lim at Chungbuk Tobacco Control Center of Chungbuk National University College of Medicine for valuable suggestions. Hyun Woo Park and Jong Seol Lee at School of Electrical and Computer Engineering, Chungbuk National University to deserve our special thanks for analyzing data and support.

REFERENCES

[1] World Health Organization, "WHO report on the global tobacco epidemic, 2013: enforcing bans on tobacco advertising, promotion and sponsorship," World Health Organization, 2013.
[2] *WHO Report on the Global Tobacco Epidemic*, 2008: The MPOWER package, World Health Organization, Geneva, 2008.
[3] M. Eriksen, J. Mackay, N. Schluger, F. Islami, and J. Drope, "The tobacco atlas," *Atlanta: The American Cancer Society*, 2015, 2017.
[4] American College of Obstetricians and Gynecologists, "Tobacco use and women's health," *Committee Opinion No. 503*, Obstet Gynecol, 2011.
[5] S. Kim, "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*, Jan 2012.
[6] R. Charafeddine, S. Demarest, I. Cleemput, H. Van Oyen, and B. Devleesschauwer, "Gender and educational differences in the association between smoking and health-related quality of life in Belgium," *Preventive medicine*, 2017.
[7] I. Khati, G. Menvielle, A. Chollet, N. Younès, B. Metadieu, and M. Melchior, "What distinguishes successful from unsuccessful tobacco smoking cessation? Data from a study of young adults (TEMPO)," *Preventive Medicine Reports*, vol. 2, pp. 679-685, 2015.
[8] Y. C. Chen, T. Suzuki, M. Suzuki, H. Takao, Y. Murayama, Y. Ohwada, and Y. Hayato, "Building a classifier of onset stroke prediction using random tree algorithm," *International Journal of Machine Learning and Computing*, vol. 7, pp. 61-66, 2017.
[9] A. Matsumoto, S. Aoki, and H. Ohwada, "Comparison of random forest and SVM for raw data in drug discovery: Prediction of radiation protection and toxicity case study," *International Journal of Machine Learning and Computing*, vol. 6, no. 2, pp. 145-148, 2016.
[10] P. Jaganathan, N. Rajkumar, and R. Kuppuchamy, "A comparative study of improved F-score with support vector machine and RBF network for breast cancer classification," *International Journal of Machine Learning and Computing*, vol. 2, no. 6, pp. 741-745, 2012.
[11] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, pp.1157-1182, March 2003,
[12] H. W. Park, E. Batbaatar, D. Li, and K. H. Ryu, "Risk factors rule mining in hypertension: Korean National Health and Nutrient Examinations Survey 2007–2014," *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2016 IEEE Conference*, IEEE, pp. 1-4, October 2016.
[13] K. H. Park, M. I. M. Ishag, K. S. Ryu, M. Li, and K. H. Ryu, "Efficient ensemble methods for classification on clear cell renal cell carcinoma clinical dataset," in *Proc. Asian Conference on Intelligent Information and Database Systems. Springer,* pp. 235-242, March 2018.
[14] V. Bewick, L. Cheek, and J. Ball, "Statistics review 14: Logistic regression," *Critical Care*, 2005, p. 112.
[15] H. Trevor, T. Robert *et al*., "The elements of statistical learning: Data mining, inference, and prediction," 2009
[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, 1995, pp. 273-297.
[17] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, 2004, pp. 199-222.
[18] K. M. Leung, "Naive bayesian classifier," Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007.

[19] P. N. Tan, "Introduction to data mining," *Pearson Education India*, 2006.
[20] L. Breiman, "Random forests," *Machine Learning*, 2001, pp. 5-32.
[21] W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations," *NESUG Proceedings: Health Care and Life Sciences*, Baltimore, Maryland, 2010.
[22] M. Vuk and T.Curk, "ROC curve, lift chart and calibration plot," *Metodoloski zvezki*, 2006.

Khishigsuren Davagdorj received the BSc degree in printing technology engineering from Mongolian University of Science and Technology, Mongolia in 2012. She has been studying toward the Ph.D degree at the School of Electrical and Computer Engineering, Chungbuk National University, South Korea, her main research focuses on artificial intelligence, big data, biomedical and data mining in business intelligence.

Seon Hwa Yu received the BS degree in Information and Statistics from Korea University, Sejong, Rep. of Korea, in 2016. She is an operating support team member of Chungbuk Tobacco Control Center. Her research interests focuses on 'Health-related factor of Chungbuk', 'Smoking cessation and Smoking risk factor'.

So Young Kim is an associate professor in Chungbuk National University Hospital since 2014. She has been vice-president of Chungbuk Tobacco Control Center. Her main research interests include health policy and management, chronic disease management, social epidemiology, disability studies, anti-smoking policy.

Pham Van Huy received the Ph.D in Computer Science from Ulsan University, South Korea, in 2015, and M.S. degree in Computer Science from University of Sciences, Ho Chi Minh City, Vietnam in 2007. Since 2015, he has been a lecturer and researcher at Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam. His main research interests include artificial intelligence, image processing, and computer vision.

Jong Hyock Park is a professor of health policy and management at Chungbuk National University College of Medicine since 2014. He has been president of Chungbuk Tobacco Control Center. His main research interests include health policy and management, social epidemiology, disability studies, anti-smoking policy and community activities.

Keun Ho Ryu received the Ph.D. degree in Computer Science and Engineering from Yonsei University, Korea, in 1988. He is also an honorary doctorate of the National University of Mongolia. He is currently a Professor at Chungbuk National University (CBNU), South Korea, and has been a leader in the Database and Bioinformatics Laboratory, South Korea, since 1986. Also, he has served the Korean Army as an ROTC. He has worked at the University of Arizona, U.S.A., as a postdoctoral Research Scientist, and also at the Electronics and Telecommunications Research Institute, South Korea, as a Senior Researcher. He was a Vice-President of the Personalized Tumor Engineering Research Center.

Professor Ryu has served on numerous program committees including as a demonstration co-chair of the VLDB, a panel and tutorial co-chair of the APWeb, and a general co-chair of the FITAT. He has published or presented more than 1,000 referred technical articles in various journals and international conferences and is the author of several books. His research interests include temporal databases, spatiotemporal databases, temporal GIS, stream data processing, knowledge-based information retrieval, data mining, biomedicine, and bioinformatics. Professor Ryu has been a member of the IEEE and the ACM since 1983.