Determining Smokers' Quitting Behavior Patterns for Multi-level Intervention of a Smoking Cessation Program

Dennis A. Martillano
College of Computer and Information Science, Malayan Colleges Laguna, Philippines
Email: dennis.martillano@upou.edu.ph

Argel Pereña Barrameda, Katherine Joy Mendoza Domondon, and John Kenneth Leuterio Rioflorido
Malayan Colleges Laguna, Philippines
Email: {argel.barrameda, katherine.domondon, johnkenneth.rioflorido}@gmail.com

***Abstract*—Smoking Cessation is the medical process where a smoker undergoes a range of procedures to quit smoking. Related studies observe that a smokers' ability to quit was dependent on patterns to a smoker's behavior. With this, the study aims to develop a model that determines quitting behavior patterns of smokers undergoing cessation. Smoking Cessation dataset was acquired from an identified municipal cessation center in the Philippines. The dataset was subjected to Classification via Clustering technique, to identify different classes/groups of quitting patterns, utilizing attributes related to smoking behavior. Results reveal four (4) distinct clusters of quitting patterns with the consideration of the Elbow Method, which then underwent proper Behavioral Pattern labeling, through the guidance of a public health expert. These labels were included as an additional attribute in the final dataset before classifying. The final model was integrated into the developed web application for Smoking Cessation Center website, which enables public health officers and medical practitioners in predicting the smokers' quitting behavior pattern.**

***Index Terms*—medical data mining, classification via clustering, elbow methodology, smoking cessation**

I. Introduction

Smoking cessation or "treatment of tobacco dependence" refers to a range of techniques including motivation, advice and guidance, counselling, telephone and internet support, and appropriate pharmaceutical aids all of which aim to encourage and help tobacco users to stop using tobacco and to avoid subsequent relapse. Cutting the smoking rate will require new interventions, enhance and effectiveness of existing cessation treatments and maximize the utilization of both [1].

A smoking cessation center in a municipality in the Philippines implements a multi-level intervention program. The program is patterned to international cessation mandate, followed by local health organizations. The first stage of the smoking cessation program happens when health workers from the center are contacted or visited by those who wish to undergo the program. The "Five A's" model is administered. The model includes: (1) Ask about smoking, (2) Advise smokers to stop, (3) Assess the smoker's willingness to stop, (4) Assist those smokers willing to stop, and (5) Arrange follow-up. With the use of the Five A's model and motivation, chances are, the smokers will cease to smoke or relapse back to smoking. If smokers relapse, they are then elevated to counselling which is the second stage of the smoking cessation program. A quitting contract between the counselor and the smoker is made in order to strengthen the commitment of the smoker to smoking cessation. Through regular counselling, the smoker is expected to cease smoking but, if the smoker relapses back to smoking, the case will be elevated to the third stage of the smoking cessation program, which is through advanced psychological and pharmacological means.

According to a paper review on the smoking patterns in the general population, behaviors relating to smoking cessation are prevalent and are one of the different factors related in the attempt to succeed in smoking. This include psychological, pharmacological and social factors that contribute to uptake and maintenance of smoking [2]. In addition, statistics show relapse of patients to smoking cessation programs. This suggests interesting gaps in smoking cessation programs, including interventions given and patients' behavior towards these, which can further be understood through a study.

A key challenge to greater progress in smoking cessation success is to directly identify the quitting behavior pattern of the patients, which could lead to direct identification of the right treatment and intervention. Interventions include medication, and replacement therapy such as gum or patch and abstinence without the help of a pharmacologic assistance. Different smoking cessation strategies can improve clinical and patient decision-making [3].

To be able to provide a model that determines quitting behavior patterns of smokers, specific objectives were laid down. First is to filter out and select applicable attributes from the preprocessed datasheet. Next is to

Manuscript received April 20, 2020; revised September 23, 2020.

doi: 10.12720/jait.11.4.200-208

develop a model that will be able to identify group of behavioral patterns using Clustering techniques, which can be used in classifying records according to cluster/group to which they belong. Third is to analyze the model and subject the behavioral patterns in the model to proper medical labeling. And lastly, to integrate the model to a web application.

The study is geared towards provision of possible new insight, where health care practitioners and officers can get information of how a smoker behaves on the process of smoking cessation that will help them provide the appropriate medication and interventions needed for successful cessation of smoking.

II. Literature Review

A. Data Mining and Smoking Cessation

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous as viewed in a paper review about medical data [4]. These data need to be collected in an organized form. One of the healthcare area today that generates large amounts of complex data about patients is smoking and smoking cessation. These large amounts of data are a key resource to be processed and analyzed for knowledge extraction [5].

Smoking is now increasingly rapidly throughout the developing world and is one of the biggest threats to current and future world health. Encouraging smoking cessation is one of the most effective and cost effective things that doctors and other health professionals can do to improve health and prolong their patients' lives [6].

Smoking cessation is a multi-dimensional behavior, related to physiological, biological, psychological, social, and community factors that may emerge in a classification model [7]. It is also one of the most cost-effective interventions available to health-care systems. Most smokers will attempt to quit on their own. However, unaided quitting is associated with the lowest change of long-term success.

B. Behavior Pattern Discovery in Smoking

Personality factors are likely to have an important role in determining who begins, continues, and successfully stops smoking. Personality differences could potentially be important in the etiology, persistence, and termination of smoking behavior [8]. The study proved the link between smoking behavior and personality.

Another study highlighted personality characteristics of the smoker as an obstacle to smoking cessation [9]. The study revealed that smokers tend to be more extroverted, anxious, tense, and impulsive, and show more traits of neuroticism and psychoticism than do ex-smokers or nonsmokers.

C. Attributes and Techniques in Behavioral Pattern Discovery

In the 1994 study previously mentioned at [8], gender, age, average cigarettes smoked per day, years spent smoking, and status indicating a smoker or an ex-smoker were used. For the participants to be considered as a smoker, they had to have smoked more than 100 cigarettes; it was to compare the personalities of a smoker and a non-smoker.

According to a study about Psychological characteristics associated with tobacco smoking behavior, that examined the relationship between nicotine dependence with smoking cessation and major depression, the more a measure of dependence is based exclusively on level of daily smoking, the greater is its ability to predict cessation [10]. They used Fagerstrom Test for Nicotine Dependence (FTND) and the Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition (DSM-III-R) on 238 daily smokers with respect to nicotine dependence.

In Malaysia and Thailand in 2005, 4,004 smokers were surveyed and 2,426 smokers were followed up in 2006. Baseline measures of sociodemographic, dependence and interest in quitting were used as attributes. Sociodemographic variables were sex, age, race, dwelling, education, and income. Other information taken were cigarettes per day, intention to quit or motivation, and outcome expectant. Results stated that smoking fewer cigarettes per day, higher levels of self-efficiency, and more immediate quitting intentions were predictive of both making a quit attempt and staying quit in both countries [11].

Another study [12] had the objective of investigating the smoking behavior and reasons for failure of smoking cessation in the general population and among health care professionals in Metro Manila, Philippines. Principal reasons for smoking and difficulty quitting in the general population are social and environmental goals, particularly in the workplace, and associative processes.

A study also aimed to examine changes in patterns of smoking initiation and cessation among vocation high school students. They concluded that social influences and self-efficacy attributes examined in this study, predict both youth smoking initiation and youth smoking cessation [13].

D. Classification via Clustering

Classification is a supervised learning method to extract models describing important data classes to predict future trends. Clustering is the process of grouping similar elements. This technique may be used as a preprocessing step before feeding the data to classify model.

The classification via clustering approach is based on the "clusters to classes" evaluation routine in the cluster evaluation code, which finds a minimum-error mapping of clusters to classes [14]. This study concluded that the classification via clustering approach obtained similar accuracy to traditional classification algorithms

III. METHODOLOGY

The study's conceptual framework is in an Input Process Output Format as shown in Fig. 1.

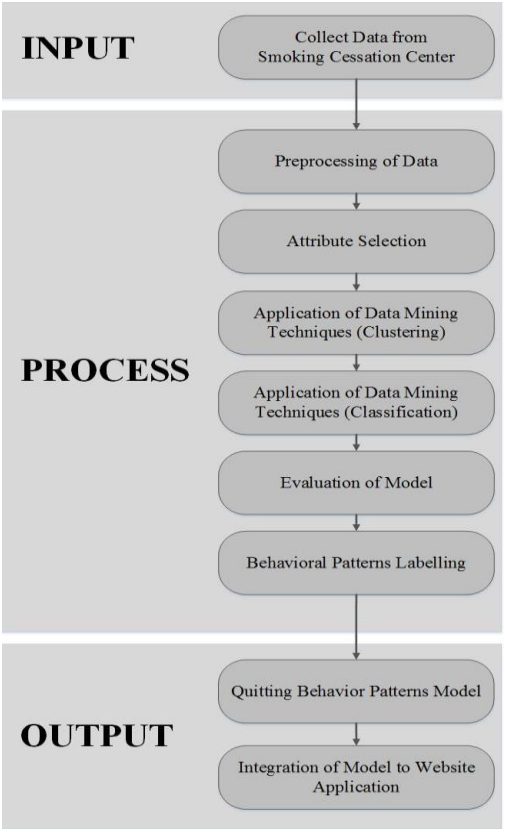

Figure 1. Conceptual framework.

Input. Data was collected with the consensus from an identified local municipal Smoking Cessation Center in the Philippines. The Smoking Cessation Center provided the dataset of smokers who underwent cessation program. Most of the records are in printed forms and logs, which are organized and piled in folders. Forms are common and structured according to local health organization mandate. Few redundant records, however, were electronically recorded as database files.

Process. The data from the cessation center was processed. Preprocessing was required to fix missing information and make it more understandable. The smokers' information was organized collated electronically in a spread sheet. Certain information that are deemed unnecessary and may slow down queries were removed. In transformation, overlays were added such as the demographic overlays, to make the data usable and navigable [15]. Under preprocessing, attribute selection was performed. The collected data was scanned for attributes, and the significant factors that affect the result of the cessation of smokers. Non-significant attributes were removed for better accuracy. After the attribute selection and complete preparation of data, the dataset underwent Clustering technique to find group of patterns and form classes that can be used for prediction. Clusters/classes created, were subjected to the behavioral patterns labelling, which was performed by the Head Counselor from the Smoking Cessation Center where they provide proper labels and descriptions. Once labelled, records in the dataset were assigned with clusters before performing Classification technique. The generated model from the Classification was subjected to evaluation through Kappa and Accuracy Values.

Output. A predictive model that determines quitting behavior pattern of smokers was generated. A web application was developed separately where model was integrated. The Website also serves as an application for data entry and collection, and as an aid for the counselors and health workers in determining the class/cluster a smoking cessation patient belongs to.

IV. RESULTS AND DISCUSSION

A. Data Gathering

The data used for this study came from an identified Municipal Smoking Cessation Center in the Philippines. Initially, the center showed a sample record for reference. The researchers were advised to return once the center had gathered all their existing records. The first set of the data consisted of 31 unfiltered document forms from the Smoking Cessation Center which covered the year 2015 until September 2016. The second set of data consisted of 124 records that came from the previous database of the Health Center's e-health system in the form of an unfiltered text file. The 31 unfiltered document forms were then encoded into an Excel spreadsheet while the 124 unfiltered text file records were merged into the same spreadsheet. The spreadsheet contained 155 records after merging the datasets.

B. Data Preparation and Attribute Selection

The dataset underwent data preprocessing which included data cleaning and attribute selection. Records with null values and incomplete records were removed from the dataset to avoid noise and unnecessary results. This was also done to ensure the quality of the dataset, the completeness of the records, the accuracy of the dataset to represent the population, and integrity of the dataset. In total, 11 records were identified as consultation records for e-health services, 3 records had names that were blank, and 31 records have incomplete entries. The incomplete records were data with no answers to the questions and no relapse records. From the original record count of 155, 45 records were removed for data cleaning.

After data cleaning, attribute selection was done. The dataset provided attributes that were in relation to behavioral patterns and personal information of the smokers. The fields acquired from the data collected were under three categories: Personal Information, Knowledge of the Ill effects of smoking and benefits from quitting, and Smoking behavior and Nicotine Dependence. Fields supported by the Related Literature mentioned prior to this section were: the type of smoker, age they started smoking, trigger that made them start smoking, years spent smoking, number of sticks consumed per day, trigger to smoke, environment, and interest to quit.

Personal Information Category is consist of name, age, gender, civil status, address, level of education, position level, employment status, and monthly salary. Name, address, position level, and monthly salary were removed for data privacy reasons. Age, gender, employment status, civil status and level of education were supported as sociodemographic dependence fields while position level and monthly salary were viewed as irrelevant.

Knowledge about the ill effects of smoking and benefits from quitting Category questions and identifies the smoker's awareness of their standing. Personality factors that are related to a smoker's knowledge and awareness show smoking cessation patients' motivation to quit.

Smoking Behavior and Nicotine dependence Category asks about the type of smoker and their reasons for smoking. According to related literature, the behavior of the smoker is important in finding the results of a smoking cessation program. The questions provided in this part allows the consultant or doctor to know the smoker and plan their cessation process. The questionnaire asks about the following: the type of smoker you are (a regular smoker, a social smoker, or an ex-smoker), the smoker thinks they are, at what age they started smoking and the trigger, when they usually smoke and why they feel the need to, their average sticks per day, at what time of the day they smoke their first stick, location where they smoke and where they buy, whether they have tried to quit smoking, the reason for quitting, number of times they tried to quit, what they think is the effective way for them to quit, and the support they think they need to successfully quit.

Fields supported by the related literature were: the type of smoker, age they started smoking, trigger that made them start smoking, years spent smoking, number of sticks consumed per day, trigger to smoke, environment, and interest to quit.

Attributes that were included and provided but not supported were: at what time of the day they smoke their first stick, location where they buy, tried to quit smoking, the reason for quitting, number of times they tried to quit, what they think is the effective way for them to quit, and the support they think they need to successfully quit.

The basis for adding the mentioned attributes was according to the recommended personality factors such as impulsiveness, anxiousness, sensation seeking factor, and motivation. "First cigarette after waking up" could be the result of impulsiveness and the sensation-seeking factor depending on the time. "Location where cigarette is bought" offers neighborhood store, offices, and store near the offices. The location may differ in stress and self-presentation concerns provided. "Number of times they tried to quit and its reason" could show the smoker's perseverance. Further information about how and why they relapsed, what process they used in order to quit and why it did not work assists the consultants. Also, "the persistence of a smoker after relapsing", "the knowledge of the ways to quit" and "the support they need" could prove the smoker's motivation.

After data preparation, from the original dataset record count of 155, 45 records were removed to avoid noise and unnecessary results in clustering. The final dataset consisted of 110 records had undergone data preparation and ready for clustering. Final list of attributes can be viewed in Table I, with legends found in Table II.

TABLE I. FINAL LIST OF ATTRIBUTES FOR CLUSTERING

Attributes	Data Type
Education	Nominal (Elementary, HS, College)
Age	Numeric
Gender	Nominal (Male, Female)
Civil Status	Nominal (Single, Married, Widdowed)
Q1(awareness to ill effects)	Nominal (Yes,No)
Q2(awareness to ill effects)	Nominal (Yes,No)
Q3(awareness to ill effects)	Nominal (Yes,No)
Q4(awareness to ill effects)	Nominal (Yes,No)
Employment Status	Nominal (Unempoyed,Employed)
Smoker Type	Nominal (Regular, Social, Ex-Smoker)
Age Started Smoking	Numeric
Triggered To Smoke	Nominal (C, F, P)
Smoke When	Nominal (H, S, B)
Smoke Because	Nominal (U, E, R, P)
Average Stick per Day	Numeric
Urge to Smoke after Waking Up (in minutes)	Numeric
Place to Smoke	Nominal (H, W, P, B, O)
Buy Cigarettes From	Nominal (N, O, S)
Tried to Quit	Nominal (Yes, No)
Quit Smoking Because	Nominal (T,Q,S,I)
Number of Times to Quit	Numeric
Support Needed	Nominal (C, M)
Relapsed	Nominal (Yes, No)

TABLE II. TABLE LEGENDS

Acronym	Meaning
Elementary, HS, College	**Elementary, High School, College**
Q1	**I know the ill effects of smoking**
Q2	**I know the dangers brought about by second hand smoking**
Q3	**I know the benefits of quitting smoking**
Q4	**I want to gain more knowledge about ill effects of smoking and benefits of quitting**
C, F, P	**Curiosity, Family Influence, Peer Pressure**
H, S, B	**Happy/Relaxed/Drinking, Sad/Angry/Temsed/Stressed, Bored**
U, E, R, P	**Used to it, Energy, Relaxes, Pleasure**
H, W, P, B, O	**Home, Office/Work, Public Places, Bars&Restaurangs, Others**
N, O, S	**Neighborhood Store, Office Canteen, Store near the office**

C. Clustering

K-means clustering technique was used to group the data into k number of clusters. For this study, WEKA tool was used to perform the process. İn order to obtain k-

value, the elbow method was used; this is the older method for determining the true number of clusters in a data set. İt is a visual method which is done by increasing the k value by 1 starting with k=2. At some value for k, the cost drops dramatically and after that, it reaches a plateau when you increase it further.

Based on the graph on Fig. 2, the k value for the dataset used in the study is k = 4. Since k = 4, there were 4 clusters formed and had a Sum Squared of Errors (SSE) of 445.169589.

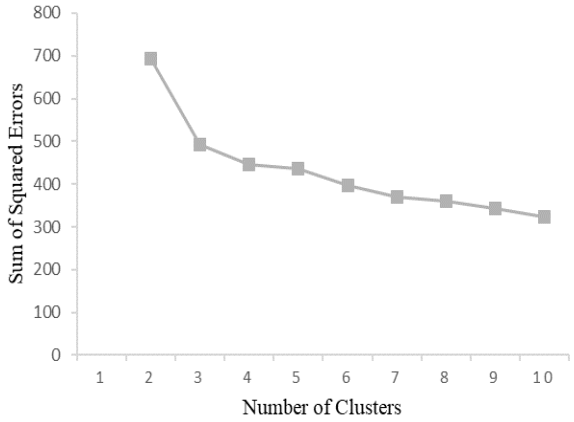

Figure 2. Result of sum of squared errors for elbow method.

The elbow method tries to find a "sweet spot" where the amount of variance explained by adding an additional cluster does not increase significantly. As the name suggests, this is "detected" by observing a change in the slope between points in a graph of the within cluster sum of quares vs number of clusters [16].

Alternatively, the elbow method can be applied to the percent variance explained by calculating the ratio of between-group variance (sum of squares) to the total variance. The variance is shown on Table III. Cluster count of 1-2 has variance of 201211.23. The 2-3 cluster count has 1096.023. Significantly, the variance of the clusters 3 and 4 drops to 50.131612. Therefor, 4 cluster count was utilized and four(4) clusters were formed in performing K-Means.

TABLE III. SSE VARIANCE FOR ELBOW METHODOLOGY

Clusters	Variance
1-2	20211.230000
2-3	1096.023000
3-4	50.131612

Clusters formed were labelled by a public health expert of the Smoking Cessation Center. The description of the clusters are shown on Table IV. Here, the age range for young or young adults are people between 15 and 24 or people in their late teens and mid twenties, while adults are 25 and above [17]. The age idicated in each cluster is the mean of all the ages from the data collected.

TABLE IV. SUMMARIZED DESCRIPTION OF CLUSTERS

Attributes	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Education	College	Elementary	High School	High School
Age	22.7857	17.5	46.2439	44.0968
Gender	F	M	M	M
Civil Status	Single	Single	Married	Married
Q1	No	Yes	Yes	No
Q2	Yes	Yes	Yes	Yes
Q3	Yes	Yes	Yes	Yes
Q4	Yes	Yes	Yes	Yes
Employment Status	Employed	Unemployed	Employed	Employed
Smoker Type	Social	Social	Regular	Regular
Age Started Smoking	14.6429	13.2917	18.2439	14.5161
Triggered To Smoke	Peer Pressure	Peer Pressure	Family Influence	Family Influence
Smoke When	Sad & Bored	Sad	Happy	Bored
Smoke Because	Smoking Relaxed Me	Gives me pleasure	Smoking Relaxed Me	Smoking Relaxed Me
Average Stick per Day	5.5714	7.25	15	22
Urge to Smoke After Waking Up (Mins)	73.5714	70.4167	10.9756	19.0323
Place to Smoke	Public Places & Bars	Public Places	Home & Work	Home
Buy Cigarettes From	Store Near Office	Neighborhood	Store Near Office	Neighborhood
Tried To Quit	Yes	Yes	Yes	Yes

Cluster 0 as shown in Table III, are predominantly young female professionals who are currenlty employed. This cluster consists of casual smokers who usually start at a young age, triggered by peer preassure, and smokers when sad and/or bored. They smoke mild to average number of cigarettes a day. It is interpreted that they are driven socially to smoke due to their environment and it's influence. These smokers mostly chooses gradual quitting with couseling but has tendency to relapse.

Cluster 1 as shown in Table III, is made up of predominantly young unemployed men who are most likely studying due to their age and level of education. They have sufficeint knowledge about smoking and are more driven socially. Most of them choose abrupt quitting and are most likely to succeed in cessation through counseling. It could be because of their impulsive behavior which can be controlled compared to compulsive ones.

Cluster 2, shown on Table III, are predominantly adult men who are regular smokers with sufficient knowledge about smoking. They are considered very compulsive since they are knowledgable but chooses not to quit, shown in the number of times they tried to quit. The compulsive behavior is affected by the social factors, since it's triggered by peer preasure and driven by social

behavior. Through gradual quitting, they choose counseling and will most likely succeed in cessation with it.

Cluster 3 in Table III, are adult regular smokers who have smoked for many years with knowledge of the ill-effects of smoking. They are also considered compulsive since they are aware but chooses not to quit, they have a small frequency of times tried to quit. They differ from Cluster 2 by not being affected by social factors. Through gradual quitting, this group chooses medication and will most likely relapse cessation.

After clustering, the described attributes were sent to the Cessation Center's Head Counselor for labeling. He validated the results and their descriptions based on the smoker's behavior, and labeled the clusters and provided their descriptions as shown in Table V.

TABLE V. LABELED CLUSTER DESCRIPTION

Clusters	Label	Description
Cluster 0	**Casual Drive**	Young professionals who smoke casually or socially. Casual Smoking. Person belonging in this cluster usually triggered by peer pressure. This behavior is driven by the fact that they are most likely socially active and have access to places where smoking among young people with professional job is prevalent.
Cluster 1	**Early Impulsive-Drive**	Predominantly very young smokers who are social smokers. Their social behavior with respect to making sudden decisions that leads them to smoking brought about by peer pressure is due to the fact they are young and vulnerable
Cluster 2	**Compulsive-Social Drive**	Smokers belonging to this type or cluster are predominantly regular male smokers whose smoking habits are domestically at work. Compulsive type is usually described as a smoker for a long period of time who is aware of the ill effects of smoking.
Cluster 3	**Compulsive Drive Type**	Most of Compulsive smokers are regular smokers who have smoked for many years. They are usually not into quitting (as seen w/ the number of times to quit) but they are aware of the ill effects of smoking.

D. Results of Classification

In order to allow future data from smokers to be classified into their respective clusters, classification was conducted on the dataset that included the assigned cluster. The dataset for clustering was appended with a new attribute which is the Assigned Cluster. The new dataset with a new appended attribute (Cluster) was subjected to the classification techniques using WEKA tool.

Based on the results of the comparison of algorithms in Table VI, Naive Bayes had the highest accuracy and kappa, followed by K-nearest neighbor, and J48. Even though Naive Bayes had the best performance, the researchers opted to use K-nearest neighbor as recommended by previous recent studies that used classification as suitable for data mining in healthcare. KNN was then chosen as the algorithm to be used.

TABLE VI. CLASSIFICATION ALGORITHM COMPARISON BASED ON KAPPA AND ACCURACY

Classification Algorithm	Kappa	Accuracy
J48	0.77	0.84
KNN	0.86	0.90
Naive Bayes	0.89	0.92

KNN also requires selection of k which represents the number of neighbors to be used in classification. The kappa and accuracy values were graphed and showed a fluctuation in the values as shown in Fig. 3 and Fig. 4. The study chose the point in which the kappa and accuracy started to stabilize. The value chosen as the number of K for KNN is 5.

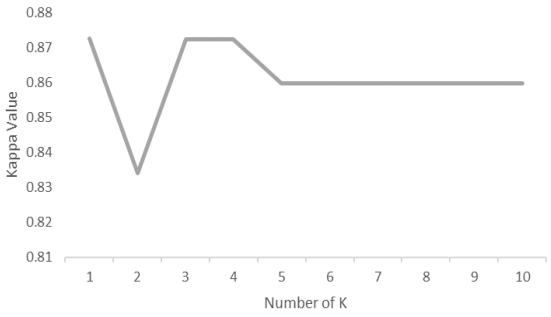

Figure 3. Kappa of k-nearest neighbor algorithm.

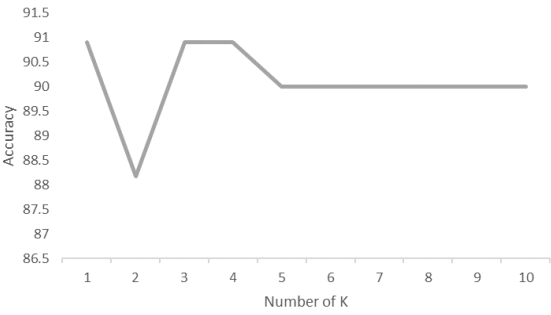

Figure 4. Accuracy of k-nearest neighbor algorithm.

The KNN model was then tested using 5 actual new test data from the center. Table VII shows the Results of KNN Model Testing with the test data. It shows how each test data falls under a certain cluster. This was decided by which cluster had the highest percentage of distribution.

TABLE VII. RESULTS OF KNN MODEL TESTING

Test Data	Model Prediction	Cluster Distribution			
		0	1	2	3
1	Cluster 1	0.2994	66.1677	33.2335	00.2994
2	Cluster 3	00.2994	33.2335	00.2994	66.1677
3	Cluster 2	00.2994	00.2994	99.1018	00.2994
4	Cluster 1	00.2994	99.1018	00.2994	00.2994
5	Cluster 0	99.1018	00.2994	00.2994	00.2994

Each new test data are evaluated after the clusters have been chosen. Table VIII shows an evaluation of the class prediction of the test data.

TABLE VIII. EVALUATION OF TEST DATA

Test Data	Evaluation
1	This falls under Cluster 1, Early Impulsive-Drive a young social smoker. Their impulsive behavior is brought by peer preasure and is due to the fact they are young and vulnerable. Test data 1 is likely to succedd in cessation through counseling. Test Data 1 falls quite close to Cluster 2 as well, Compulsive-Social Drive, smokers that fall to this cluster are predominantly regular male smokers.
2	Test Data 2 falls under Cluster 3, Compulsive Drive Type of smoker. Most of these smokers are regular smokers who have smoked for many years. As seen with number of times they've tried to quit, they are usually end up relapsing but are aware of the ill effects of smoking. Though test data 2 leans more on Cluster 3, 33% leans on Cluster 1, Early Impulsive-Drive. Their social behaviors make them vulnerable to their enviroment and smoking is brought about by peer pressure.
3	This data set highly falls under Cluster 2, Compulsive-Social Drive. Smokers belonging to this type of cluster are regular male smokers who usually smokes at work. They are described as a smoker who have smoked for a long period of time but is aware of its ill effects.
4	This data set highly falls under Cluster 1, Early Impulsive-Drive. A young social smoker, their social behavior with making sudden decisions leads them to smoking brought by peer pressure. Most of them usually have access to places where there are other smokers or cigarettes, making them with their young age vulnerable.
5	This data set highly falls under Cluster 0 which are young professionals who smoke casually or socially. Persons who belong to this cluster are usually triggered by peer pressure, are socially active, and have access to places where smoking among young people with professional job is prevalent.

E. Smoking Cessation Center Website and Model Integration

The last objective of the study is to integrate the model to a Web Application for the use of the Smoking Cessation Center. The model from the classification via clustering was integrated into the website for the Smoking Cessation Center.

The integration of the clustering model to the website is done by calling a jar file that resides in a folder where the website files can be found. A dedicated folder named KNN_Classifier was set to contain the required files to perform clustering and classification.The files are as follows: (1) KNN_Classifier.jar which classifies the KNN to be implemented, (2) kNN_model.model which contains the clustering model for KNN, and (3) WEKA.jar, an embedded WEKA jar file for clustering and classification. Before proceeding with the classification, the website needs to save a csv file which contains patient data record. After saving the csv file, a PHP file will call the jar file and perform the classification process. The jar file will delete the patient data input. A file named Result.txt will be placed in the same folder, and be accessed by the website to display the result through the user interface.

Fig. 5 and Fig. 6, are sample screenshots of the Smoking Cessation website.

Figure 5. Screenshot of the smoker data input page of the website.

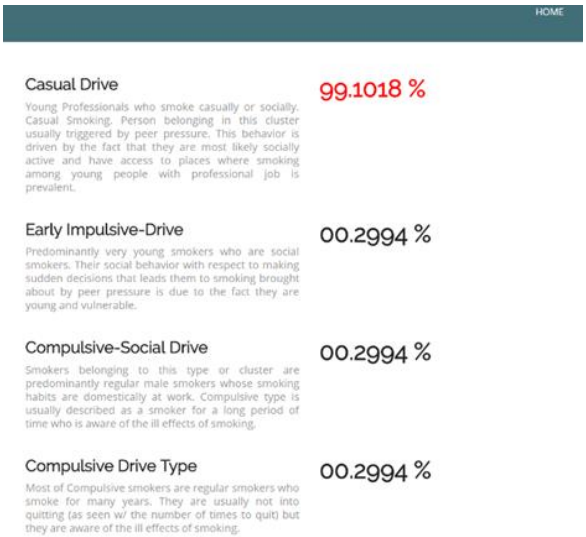

Figure 6. Screenshot of the sample smoker data interpretation.

The Web Application was then presented and rated by the Cessation Center's Head Counselor. A Usability Test (UT) using four-point Likert scale was used as an instrument to rate the website's General Appearance, Navigation, and Usefulness as a predictive tool. For the General Appearance, the total mean was 4.5, interpreted as "Strongly Agree" in accordance to design, the Ease of Use/Navigation was also evaluated with a total mean of 4.25 which has a "Strongly Agree" verbal interpretation. Finally, for the Usefulness Attributes Evaluation, the total mean was 4 with a verbal interpretation of "Agree".

V. CONCLUSION

In this paper, a model that classifies and determines the smoker's quitting behaviour pattern was developed. Through the developed model, public health officers in a smoking cessation center can be presented with descriptive information through percentile chances of a smoking cessation patient being in a specific class of quitting behaviour pattern. These information can be used as inputs in giving the proper treatment for different classes or types of smokers.

Filtering out and selecting attributes from the pre-processed datasheet was carefully done by understanding variables that are related to quitting behaviour patterns of a smoker. This was done with the use of related literatures as guide, which are solidified by the data from the actual smoking cessation center and public health expert. Clustering was performed to generate interesting classes/groups that show quitting patterns of smokers. With the help of Elbow Method in determining number of classes to generate, the clustering procedure successfully identified four (4) clusters, resulting to four types/classes of smokers. These clusters were analysed and subjected the behavioural patterns described in each cluster to proper labelling with the assistance of public health officer. The clusters were labelled as *Casual Drive*, *Early Impulsive Drive*, *Compulsive-Social Drive, and Compulsive Drive Type*. These labels helped in executing Classification via Clustering, which provides a more understandable predictive results when the model was integrated in a web application. Positive results have manifested in testing the model using actual new test data in the cessation model. In fact, Head counselors in the smoking cessation center were able to assess five (5) actual cases of smokers using the model with better interpretation of smokers quitting patterns. This is supported by an "Agree' results of the UT conducted concerning the functionalities and usefulness of a web app.

However, the amount of data is a redeeming factor to any data mining study thus, it is recommended to feed more data into the Web Application. The study advised the Cessation Center to continue adding data to the Web Application's storage in order to provide bigger dataset for future improvement of a behavioral pattern model. It would also better to use the web app as the facility to commence adding data from different cessation centers with the same type of forms, and with similar attributes.

The study also does not directly aim to replace the functionalities of the public health officers, but rather use results of the model to aid and assist the Smoking Cessation Center in assessing proper interventions. Most of the Smoking Cessation Centers are dependent on a framework when it comes to type of intervention given to the smoker. Using the Web Application on different Centers with the same framework is highly suggested. It is also suggested making the model dynamic to continue analyzing data as it is fed, thus, making the model's results more reliable with no further coding required.

The created model was integrated to a locally hosted Web Application within the cessation center. Knowing that there are some smokers who may have no time to go to the center, or receive counseling directly, its availability could reach the actual smokers in their place of leisure. Making the Web Application online could however affect the Cessation Center's and Smoker's information security. It is further recommended to make the Web Application online while adding security and additional features for maintenance.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

D. A. Martillano is the main author and responsible for the overall analysis and evaluation of procedures and the results. A. P. Barrameda is in charge of major computations in the Clustering and Classification. J. K. L Rioflorido performed the data gathering and cleansing, and K. J. M. Domondon summarized results and collected data.

ACKNOWLEDGMENT

This work was supported by College of Computer and Information Science Department of Malayan Colleges Laguna, Pulo Diezmo, Cabuyao Laguna, 4025 Philippines.

We are deeply grateful and thankful to the Public Health officers in a municipal Smoking Cessation in the Philippines for the data and technical guidance.

REFERENCES

[1] N. K. Cobb, A. L. Graham, M. J. Byron, and D. B. Abrams, "Online social networks and smoking cessation: A scientific research agenda," *Journal of Medical Internet Research*, vol. 13, no. 4, p. e119, 2011.

[2] R. West, "Tobacco smoking: Health impact, prevalence, correlates and interventions," *Psychol Health*, vol. 32, no. 8, pp. 1018-1036, 2017.

[3] M. E. Mayorga, O. S. Reifsnider, S. B. Wheeler, and R. E. Kohler, "A discrete event simulation model to estimate population level health and economic impacts of smoking cessation interventions," in *Proc. the Winter Simulation Conference*, December 2014, pp. 1257-1268.

[4] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43-48, 2011.

[5] S. Islam, M. Hasan, X. Wang, H. D. Germack, and Noor-E-Alam, "A Systematic review on healthcare analytics: Application and theoretical perspective of data mining," *Healthcare*, vol. 6, no. 2, p. 54, 2018.

[6] P. Desikan, K. W. Hsu, and J. Srivastava, *Data Mining for Healthcare Management*, 2011

[7] R. Edwards, "The problem of tobacco smoking," *Bmj*, vol. 328, pp. 217-219, 2004.

[8] M. R. Poynton and A. M. McDaniel, "Classification of smoking cessation status with a backpropagation neural network," *Journal of Biomedical Informatics*, vol. 39, no. 6, pp. 680-686, 2006.

[9] I. M. Lipkus, J. C. Barefoot, R. B. Williams, and I. C. Siegler, "Personality measures as predictors of smoking initiation and cessation in the UNC Alumni Heart Study," *Health Psychology*, vol. 13, no. 2, p. 149, 1994.

[10] R. D. C. Rondina, R. Gorayeb, and C. Botelho, "Psychological characteristics associated with tobacco smoking behavior," *Jornal Brasileiro de Pneumologia*, vol. 33, 2007.

[11] N. Breslau and E. O. Johnson, "Predicting smoking cessation and major depression in nicotine-dependent smokers," *American Journal of Public Health*, vol. 90, no. 7, p. 1122, 2000.

[12] L. Li, *et al.*, "Predictors of smoking cessation among adult smokers in Malaysia and Thailand: Findings from the International Tobacco Control Southeast Asia Survey," *Nicotine & Tobacco Research*, vol. 12, suppl. 1, pp. S34-S44, 2010.

[13] M. A. L. Tan and G. Dy-Agra, "Smoking behavior and practices and smoking cessation in the general population and among health care professionals in Metro Manila," *Phil. J. Internal Medicine*, vol. 47, pp. 129-135, 2009.

[14] F. C. Chang, *et al.*, "Social influences and self-efficacy as predictors of youth smoking initiation and cessation: A 3-year longitudinal study of vocational high school students in Taiwan," *Addiction*, vol. 101, no. 11, pp. 1645-1655, 2006.
[15] M. I. Lopez, J. M. Luna, C. Romero, and S. Ventura, "Classification via clustering for predicting final marks based on student participation in forums," International Educational Data Mining Society, 2012.
[16] J. Hsu, "Data mining and business intelligence: Tools, technologies, and applications," in *Data Mining and Business Intelligence: Tools, Technologies. Business Intelligence in the Digital Economy: Opportunities, Limitations and Risks: Opportunities, Limitations and Risks*, IGI Global, 2003, p. 141.
[17] A. Brooks. scalefreegan.github.io. [Online]. Available: http://scalefreegan.github.io/Teaching/DataIntegration/practicals/p3.html

Dennis A. Martillano is a graduate of Bachelor of Science in Computer Engineering. He has units in Interdisciplinary Studies from the University of the Philippines-Diliman. He is a Master's degree holder in Information Technology and has a Graduate Degree in Distance Education from the University of the Philippines. He is currently taking up his Doctor in Information Technology where he is specializing in Medical Data Mining in Technological Institute of the Philippines. Currently, he is teaching in the College of Computer and Information Science at Malayan Colleges Laguna, Philippines. Among his research outputs are in the area of Machine Learning, Internet of Things, Mobile Application, and Embedded Systems.

Argel P. Barrameda was born on the 27th of October in the year 1993. He is currently attending his last year in Malayan Colleges Laguna with a degree program in Computer Science with specialization in CISCO Network Administration. His curiosity and love of technology was the reason behind taking a degree in the field of Computer Science. One of his life-long goals is to contribute to the field of computing for he believes that as a part of the community, it is one's goal to not only study but also to contribute in return and add more to the extensive body of knowledge of computing.

Katherine Joy M. Domondon is a fourth year Computer Science student in Malayan Colleges Laguna. She was born on March 24, 1994 and her current residence is in Calamba, Laguna. Her specialization is CISCO Network Administration track. She chose the Computer Science track because of her curiosity in technology and interest of how fast it changes. She is fond of different kinds of creative practices and logical challenges, combining these two. She finds research and programming as ideal tools in further knowing of how far technology and creativity can be merged. She hopes one day to use her skills in helping others.

John Kenneth Rioflorido is a student of Malayan Colleges Laguna with the program of B.S in Computer Science with specialization of Network Administration under the College of Computer and Information Science. He is a risk taker, a great listener and one who loves to try new things. He is a curious person. When researching, he explores the web to gain all the possible knowledge he can acquire to have a better understanding. He wants to be productive and someday become master of his field to reach his goals. One of his goals was to able to contribute in the field of Computer Science and to make a big change in the technology.