Classification of Smoking Status: The Case of Turkey

Zeynep D. U. Durmuşoğlu
Department of Industrial Engineering
Gaziantep University
Gaziantep, Turkey
unutmaz@gantep.edu.tr

Pınar Kocabey Çiftçi
Department of Industrial Engineering
Gaziantep University
Gaziantep, Turkey
pinar_kocabey@gantep.edu.tr

***Abstract*—Smoking tobacco products has become a deadly and prevalent habit. It is a known fact that smoking negatively affects the health, economic expenditures, and social life of not only users but also second hand smokers (people's exposure to smoke). Inhaling tobacco smoke can make people vulnerable to nicotine addiction. Correspondingly, second hand smokers may become daily or less than daily smokers in time. The main objective of the presented paper was to classify smoking status of people considering second-hand smoking associated attributes: allowance to second hand smoking at home, allowance to second hand smoking at work, exposure to smoke at work in the past 30 days, beliefs that second hand smoking causes serious diseases, and gender. The classes of smoking status were defined as daily smoker, less than daily smoker and no smoker (not at all). The classification was performed using multilayer perceptron that is a well known neural network approach. The results showed that multilayer perceptron could correctly classify the smoking status of people over % 68 using five attributes. The systems like tobacco use consist of many different interacting factors that make them more complex and dynamic to analyze (classify). For that reason, the % 68 accuracy level can be interpreted as sufficient to the analysis performed for this kind of problems.**

Keywords—data mining; multilayer perceptron; smoking status; second hand smoking.

I. INTRODUCTION

Tobacco smoking is one of the most widespread addictive habits that influence the behavior of people for over four centuries [1]. Smoking and chewing tobacco that had its origins from America, had spread across Europe and all over the world after the discovery of America continent by Europeans [2].

From its first use to now, it has been used by billions of people. The World Health Organization's statistics showed that although tobacco smoking kills up to half of its users, there are still one billion smokers in the world [3]. Many people also suffer from tobacco-related serious diseases such as lung cancer, heart diseases, stroke, chronic obstructive pulmonary disease, and other forms of cancer [4].

It is a known fact that tobacco smoking does not only damages its users but also negatively affects the health, economic expenditures, and social life of people who inhale the smoke of tobacco products (second hand smokers). It is responsible from 600 000 deaths those are the results of second hand smoking each year [3]. Correspondingly, even if you do not smoke, exposure to smoke of tobacco products can result in tobacco-related diseases and even deaths. For that reason, the consequences of second-hand smoking should not be underestimated.

In the light of the information provided above, the main objective of the presented paper was to analyze the capability of classification of people's smoking status using attributes those are mostly relevant to second hand smoking. In the content of this study, only five attributes were examined. Those are:

- Gender
- Allowance to second hand smoking at home
- Allowance to second hand smoking at work
- Exposure to smoke at work in the past 30 days
- Beliefs that the second hand smoking can cause serious diseases.

The classification was performed on the Global Adult Survey (GATS) Turkey data using a neural network method that is generally used for classification in data mining.

Data mining is a powerful tool that can convert the raw data into an understandable and actionable form to help to predict future trends or provide meaning to historical events [5].

In the literature, data mining has been used by several disciplines such as business, engineering, economy, and etc. The health relevant topics have also been examined using different data mining methods. For example; Kartelj [6] made an analysis using different data mining methods to classify smoking cessation status in 2010. In 2013, Huang et al. [7] performed an association rule mining analysis using a large primary care database for smoking cessation medications. In addition, Hafeth et al. [8] used text analysis to classify smoking status from user-generated content such as online forum discussions for health-care applications. Different from the literature, the presented paper explored the classification capability of smoking status of people using five different attributes.

The remaining sections of this paper were organized as follows. First, data collection and applied method were discussed in the methodology section. Subsequently, the findings of the analysis were presented in the results section. Finally, conclusions and discussions were provided at the last section.

II. Methodology

The presented paper mainly stressed on the classification of smoking status of Turkish people using different attributes via multilayer perceptron. The details about the data collection process and the used technique are provided in the next subsections consecutively.

A. *Data Collection*

In the content of this study, five different attributes were chosen for classification analysis. The required data for the analysis was obtained from the GATS Turkey. The GATS is a national household survey that aims to collect data about tobacco use and key tobacco control measures for adults aged 15 years and older [9]. It has been one of the most important tobacco use and control related surveys of the world. It has also been supported by the World Health Organization.

After the signature of the Framework Convention on Tobacco Control (FCTC) by Turkish Government in 2004, Turkey started to implement this survey for Turkish people. It has been performed two times for Turkey until now. The first one was in 2008 while the second one was in 2012. In this study, "the GATS Turkey 2012" data was used for the classification analysis. The GATS includes many questions about demographic characteristics, tobacco use status, second-hand smoking, beliefs about tobacco control measures of participants, and etc. This study mostly focused on the generic second-hand smoking questions for attributes.

A total of 9851 individuals completed the GATS 2012 in Turkey. However, only 2675 of them completed the related questions (corresponding to the attributes that was under interest of this study). That is why, 2675 instances was used for the analysis. The classes of the attributes and the smoking status of the people were given below.

- Gender: male, female
- Allowance to second hand smoking at home: allowed, not allowed but exceptions, never allowed, no rules.
- Allowance to second hand smoking at work: allowed anywhere, allowed only in some indoor areas, not allowed in any indoor areas, there is no policy, do not know.
- Exposure to second hand smoking at work in the past 30 days: yes, no, do not know.
- Beliefs that the second hand smoking causes serious diseases: yes, no, do not know.
- Smoking status of people: daily smokers, less than daily smokers, not at all (no smoker).

In this study, the answer of the question "how successful can multilayer perceptron classify the smoking status of Turkish people using only five attributes?" was examined.

B. *Multilayer Perceptron*

Data mining can be defined as the process of automatically discovering previously unknown, non-explicit, and helpful knowledge from large databases [5]. In other words, it is a useful way to find hidden patterns and relationships in data with an emphasis on large volumes of data [10]. Its major tasks require rapid and correct partitioning of huge datasets that may come with several of attributes [11]. It mainly consists of the processes: classification, clustering, and associations [6].

Classification is the most prevalent data mining methods that employs a set of previously classified instances to develop a model which can classify the targeted records [12]. Data classification is a two step processes: learning step in which a classification model is built and classification step in which the constructed model is used to predict class of a given instance [15]. Classification generally uses neural network or decision tree based algorithms [12]. In this study, multilayer perceptron that is a kind of neural network was used to classification of smoking status of people.

Neural networks consist of a set of connected input-output units where each relationship has a weight associated with it [13]. There are various kinds of neural network algorithms such as feed-forward networks and feedback networks. Multilayer perceptron is a feed-forward network and also known as a back-propagation algorithm [14].

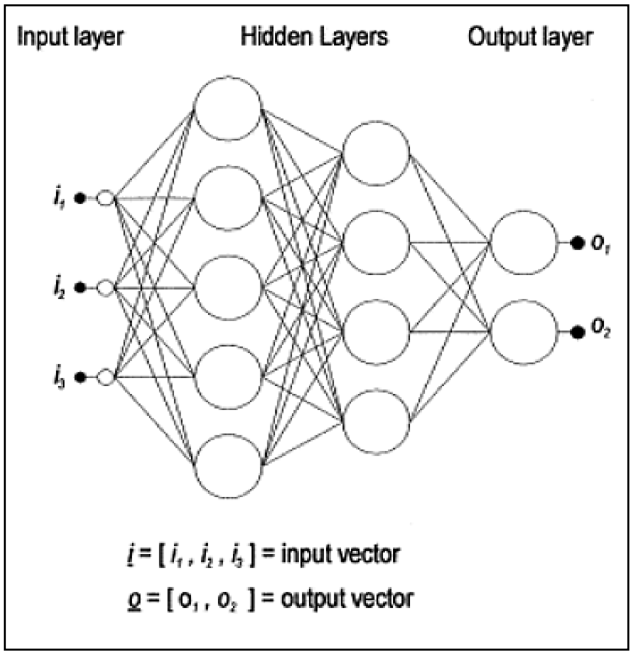
Input layer
Hidden Layers
Output layer
i_1
i_2
i_3
o_1
o_2
$\underline{i} = [\, i_1 , i_2 , i_3 \,]$ = input vector
$\underline{o} = [\, o_1 , o_2 \,]$ = output vector

Fig. 1. A multilayer perceptron with two hidden layers [14].

Figure 1 represents an example for the multilayer perceptron. The main principle of this method is that the network nodes perform calculations in the successive layers until an output value is reached after the data is presented in the input layer [15].

Waikato Environment for Knowledge Analysis (WEKA) version 3.6.13 was used to perform classification analysis with multilayer perceptron. WEKA is open source data mining software that presents a rich set of powerful machine learning algorithms for data mining tasks [16]. The results obtained using WEKA software was provided in the next section.

III. RESULTS

The analysis for the classification of smoking status of Turkish people was performed using multilayer perceptron considering five previously defined attributes. The required data was borrowed from GATS Turkey 2012. Table I represents the descriptive statistics of the studied data.

A total of 2675 instances (attributes of 2675 Turkish people) were used in the applied analysis. %75.77 of them (2027 of all instances) was male while %24.23 of them (648 of all instances) was female.

TABLE I. DESCRIPTIVE STATISTICS

Attributes	**Count** **(number)**	**Percentage** **(%)**
Overall	2675	100
Gender		
Male	2027	75.77
Female	648	24.23
Allowance to second hand smoking at home		
Allowed	554	20.71
Not allowed but exceptions	554	20.71
Never allowed	1450	54.20
No rules	117	4.37
Allowance to second hand smoking at work		
Allowed anywhere	179	6.70
Allowed only in some indoor areas	226	8.45
Not allowed in any indoor areas	2194	82.01
No policy	71	2.65
Do not know	5	0.18
Exposure to smoke at work in the past 30 days		
Yes	454	16.97
No	2210	82.61
Do not know	11	0.41
Beliefs that second hand smoking can cause serious diseases		
Yes	2579	96.41
No	78	2.91
Do not know	18	0.67
Smoking Status		
Daily	1019	38.09
Less than daily	105	3.92
Not at all	1551	57.98

There were three pre-defined classes for the smoking status of people. More than % 50 of the studied sample (%57.98 of all) was no smokers while %38.09 of them sample was daily smokers. The small amount of people (%3.92) categorized themselves as less than daily smokers.

The descriptive statistics of the examined attributes were also searched in detail. More than half of the sample (%54.20) declared that they lived at homes where smoking was never allowed. On the other hand, %20.71 of them lived in places where smoking was allowed while the same percentage of people lived in the places in which smoking was not allowed but there were some exceptions. The remaining sample said that they did not have any policy about allowance to second hand smoking at home.

Allowance to second hand smoking at work was also discussed in this study. The largest amount of the people (%82.01 of all) implied that they worked in places where smoking was not allowed in any indoor areas. However, %8.45 and %6.70 of them worked in places in which smoking was allowed in some indoor areas and completely allowed respectively. %2.65 of all did not have any policy for second hand smoking at work while %0.18 of them did not know the allowance status of second hand smoking at their works.

The statistics of the studied sample indicated that 2210 of all (%82.61) did not exposure to smoke of tobacco products at work in the past 30 days while 454 of them (%16.97) declared that they inhaled tobacco smoke at work in the past 30 days.

Lastly, the beliefs of people about the harms of second hand smoking were examined in the GATS. The vast majority of the studied sample said that they believed that second hand smoking causes serious diseases while %2.91 of them did not believe the harms of second hand smoking.

Table II represents the brief summary of the performed classification analysis. Since K-fold cross validation procedure can help the better use of the studied data [17], 10-fold cross validation procedure was also applied in this study.

A seen in table II, %68.3364 of the 2675 instances (1828 of all instances) were correctly classified by multilayer perceptron using five different attributes of Turkish people while %31.6636 of them (847 of all) were incorrectly classified.

TABLE II. RESULT SUMMARY TABLE

	Count **(number)**	**Percentage** **(%)**
Total number of analyzed instances	2675	100
Correctly classified instances	1828	68.3364
Incorrectly classified instances	847	31.6636

The details of the results obtained from classification analysis were given in table III. The values of performance measures such as precision, recall, F-measure, and ROC area can be found for each class and weighted average of them in that table.

Precision can be defined as the confidence of the analysis and calculated using the formulation given in equation (1) while recall can be defined as the sensitivity of the analysis and calculated by the formulation given in equation (2). On the other hand, F-measure is the harmonic mean of the calculated recall and precision as seen in the equation (3) whereas ROC area provides the area under receiver operating characteristics (ROC) curve.

Precision (confidence) [18] = TP/(TP+FP) (1)

Recall (sensitivity) [18] = TP/(TP+FN) (2)

F-measure [18] = 2*(Precision*Recall)/(Precision+Recall) (3)

Where TP represents true positive instances, FP and FN represents false positive and false negative instances respectively.

As seen table III, the weighted average of the confidence was 0.667 while the weighted average of the sensitivity was 0.683. The weighted average of the measures can be calculated considering the number of instances with that particular class label. The confidence of the model to assign the instances for daily smoker class was the highest (0.727) among three per-defined classes. The class of no smokers followed the daily smoker class with 0.672 confidence value.

The second class (less than daily class) had the zero value. The model was not confident to assign the less than daily smokers in the accurate class. The main reason of it can be that the number of instances related to less than daily smoker class was relatively lower than the other two pre-defined classes. The number of instances and the distribution of them to the classes are really vital for data mining analysis. In this study, the number of instances may not be enough to make confident classification for the less than daily smokers.

The sensitivity values of the performed analysis were also calculated. The weighted average of the recall was 0.683.The model has the highest sensitivity value for the no smoker class. It was around 0.92 while the sensitivity value of the daily smoker class was 0.4. Similar to the recall values, F-measure was higher for no smoker class. It was more than 0.77.

The findings of the performed analysis were given above. In general, the weighted averages of the all performance measures were more than 0.64. The correctly classified instances were also close 0.7. Even though it was not enough to make certain judgment such as "these five attributes were entirely enough to classify the smoking status of people using multilayer perceptron", it can be acceptable when the complexity and dynamisms of the tobacco use and control environment was taken into account.

TABLE III. DETAILED ACCURACY TABLE BY CLASS

Performance Measures	**Smoking Status Classes**			**Weighted Average**
	Daily smokers	Less than daily smokers	Not at all	
Precision	0.727	0	0.672	0.667
Recall	0.4	0	0.916	0.683
F-Measure	0.516	0	0.775	0.646
ROC Area	0.723	0.539	0.719	0.713

When the examined systems or problems show complex system structure, the moderate accuracy levels can also be helpful to have opinion about the studied topic. Medical problems can be example of the mentioned situations. The moderate accuracy levels can be used for the pre-diagnosis process and etc.

IV. CONCLUSION

This study was conducted to analyze the capability of the classification of smoking status using only five attributes. Gender, allowance to second hand smoking at home, allowance to second hand smoking at work, exposure to smoke at work in the past 30 days, beliefs about second hand smoking causes serious diseases were defined as the attributes to use for the classification analysis. Multilayer perceptron was used in order t o perform the analysis. The required data was retrieved from the GATS 2012 of Turkey.

A total of 2675 instances were used in this study. 1019 of these instances was previously defined as daily smokers while 1551 of them were no smokers and 105 were less than daily smokers.

The findings of the study indicated that multilayer perceptron assigned 1828 instances among 2675 (%68.33 of all) to the correct class using only five defined attributes. Even though this accuracy level can be interpreted as not so high, it can be enough when the complexity of the problem considered. The systems like tobacco use consist of many different interacting factors that make them more complex and dynamic to analyze (classify). For that reason, the moderate accuracy level can be interpreted as sufficient to the analysis performed for this kind of problems. Thus, the results of the performed classification analysis can be useful when an initial opinion is required for future studies. This can be used as an initial step for advance studies.

REFERENCES

[1] S. M. Moustafa and A. H. El-elemi, “Evaluation of probable specific immunotoxic effects of cigarette smoking in smokers”, Egypt. J. Forensic Sci., vol. 3, pp. 48–52, Jun. 2013.

[2] “A brief history of smoking”, Cancer Council NSW, 20-Dec-2011]. Available at: http://www.cancercouncil.com.au/31899/uncategorized/a-brief-history-of-smoking/. [Accessed: 14-jun-2016].

[3] “WHO | Tobacco”, WHO. Available at: http://www.who.int/mediacentre/factsheets/fs339/en/. [Accessed: 08-Sept-2015].

[4] R. E. Harris, Epidemiology of Chronic Disease. Jones & Bartlett Publishers, 2013.

[5] B. Ozisikyilmaz, R. Narayanan, J. Zambreno, G. Memik, and A. Choudhary, “An Architectural Characterization Study of Data Mining and Bioinformatics Workloads”, IEEE International Symposium on Workload Characterization, 2006, pp. 61–70.

[6] A. Kartelj, “Classification of Smoking Cessation Status Using Various Data Mining Methods”, Math. Balk. New Ser., vol. 24, pp. 199–205, 2010.

[7] Y. Huang, J. Britton, R. Hubbard, and S. Lewis, “Who receives prescriptions for smoking cessation medications? An association rule mining analysis using a large primary care database”, Tob. Control, vol. tobaccocontrol-2011-050124, Jan. 2012.

[8] “Text Analysis of User-Generated Contents for Health-care Applications - Case Study on Smoking Status Classification”. Available at: http://www.academia.edu/20328163/Text_Analysis_of_User-Generated_Contents_for_Health-care_Applications_-_Case_Study_on_Smoking_Status_Classification. [Accessed: 21-Mar-2016].

[9] “WHO | GATS (Global Adult Tobacco Survey)”, WHO. Available at: http://www.who.int/tobacco/surveillance/gats/en/. [Accessed: 15-Mar-2016].

[10] J. H. Friedman, “Data mining and statistics: What’s the connection”, Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics, 1997.

[11] A. Abraham, S. Das, and S. Roy, “Swarm Intelligence Algorithms for Data Clustering”, pp. 279–313, 2008.

[12] B. K. Baradwaj and S. Pal, “Mining Educational Data to Analyze Students’ Performance”, ArXiv12013417 Cs, Jan. 2012.

[13] J. Han, J. Pei, andM. Kamber, Data Mining: Concepts and Techniques. Elsevier, 2011.

[14] M. W. Gardner and S. R. Dorling, “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences”, Atmos. Environ., vol. 32, pp. 2627–2636, Agu. 1998.

[15] H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, “A multilayer perceptron-based medical decision support system for heart disease diagnosis”, Expert Syst. Appl., vol. 30, pp. 272–281, Feb. 2006.

[16] Z. Markov and I. Russell, “An Introduction to the WEKA Data Mining System”, Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, New York, NY, USA, 2006, pp. 367–368.

[17] J. L. M. Amaral, A. J. Lopes, J. M. Jansen, A. C. D. Faria, and P. L. Melo, “An improved method of early diagnosis of smoking-induced respiratory changes using machine learning algorithms”, Comput. Methods Programs Biomed., vol. 112, pp. 441–454, Dec. 2013.

[18] “WekaDataAnalysis.pdf”. Available at: http://www.cs.usfca.edu/~pfrancislyon/courses/640fall2015/WekaDataAnalysis.pdf. [Accessed: 19-Mar-2016].