Turk J Public Health
Available online at: http://dergipark.org.tr/tjph/
doi:10.20518/tjph.884692

TURKISH JOURNAL OF PUBLIC HEALTH

ORIGINAL ARTICLE / ORİJİNAL MAKALE

Socio-demographic determinants of smoking: A data mining analysis of the Global Adult Tobacco Surveys

Sigara kullanımının sosyo-demografik belirleyicileri: Küresel Yetişkin Tütün Araştırmaları üzerine bir veri madenciliği analizi

Zeynep Didem Unutmaz Durmuşoğlu[a], Pınar Kocabey Çiftçi[b],

[a] Associate Prof. Dr., Department of Industrial Engineering, Gaziantep University, Gaziantep, Turkey.
[b] Research Assist Dr., Department of Industrial Engineering, Gaziantep University, Gaziantep, Turkey.

Received: 22.02.2021 , Accepted: 01.10.2021

ABSTRACT

Objective: This paper presented a) how the Global Adult Tobacco Surveys (GATSs) data can be used for extracting valuable information about tobacco use behaviors of people and b) the prediction performance of the implemented classification algorithms on the GATS data. **Methods:** Three well-known classification methods: K-nearest neighbor, C4.5 algorithm, and multilayer perceptron were applied to assess the classifying performance for the smoking status of GATS participants (pre-defined classes: smoker and no smoker) based on the socio-demographic characteristics (age group, gender, residence, education level, and working status). The first analysis was performed on the GATS data from Turkey. Subsequently, the model producing the best performance for Turkey was also implemented for other six European countries: Greece, Kazakhstan, Poland, Romania, Russia, and Ukraine. **Results:** All of the tree algorithms were more confident to classify no smokers. The correct classification rate of C4.5 algorithm was the highest among the algorithms for the GATS Turkey data. In addition, the C4.5 algorithm classified the males more detailed than the females. The comparative analysis indicated that the C4.5 algorithm correctly classified the smoking status of participants of Ukraine over 80% while it was lower than 70% for Greece. Thus, the effects of demographic factors on smoking status can change from one country to another. **Conclusion:** This paper indicated that the data supplied by GATS such as demographic data may help to compute the likelihood of an individual to be a smoker in the future.

Keywords: Smoking, tobacco use, public health

Correspondence: Zeynep Didem UNUTMAZ DURMUŞOĞLU, Department of Industrial Engineering, Gaziantep University, Gaziantep, Turkey. **E-mail:** unutmaz@gantep.edu.tr. **Tel:** +90 342 317 2614.

Cite This Article: Durmuşoğlu U ZD, Çiftçi K P. Socio-demographic determinants of smoking: A data mining analysis of the Global Adult Tobacco Surveys. Turk J Public Health 2021;19(3):251-262.

ÖZ

Amaç: Bu makale a) Küresel Yetişkin Tütün Araştırması (KYTA) verilerinin tütün kullanım davranışları hakkındaki değerli bilgileri ortaya çıkarmada nasıl kullanılabileceğini ve b)KYTA verileri üzerinde uygulanan sınıflandırma algoritmalarının performanslarını sunmaktadır. **Yöntem:** Üç iyi bilinen sınıflandırma yöntemi olan K -en yakın komşu algoritması, C4.5 algoritması ve çok katmanlı algılayıcısı KYTA katılımcılarının sosyo-demografik özellikleri (yaş grubu, cinsiyet, yerleşim yeri, eğitim düzeyi ve çalışma durumu) temel alınarak, sigara içme durumunu (önceden tanımlanmış sınıflar: sigara içen ve içmeyen) doğru sınıflandırma performansı değerlendirilmiştir. İlk analiz KYTA Türkiye verileri üzerinde gerçekleştirilmiştir. Daha sonra Türkiye için en iyi performansı üreten model altı farklı Avrupa ülkesi: Yunanistan, Kazakistan, Polonya, Romanya, Rusya ve Ukrayna verileri için de uygulanmıştır. **Bulgular:** Bütün ağaç algoritmaları sigara içmeyenleri tespit etmekte daha doğru sonuçlar vermektedir. C4.5 algoritmasının doğru sınıflandırma oranı, Türkiye için en yüksek olandır. Ülkeler için yapılan karşılaştırmalı analiz, C4.5 algoritmasının Ukrayna'daki katılımcıların sigara içme durumunu %80'in üzerinde doğru bir şekilde sınıflandırabildiğini ancak Yunanistan için bu oranını %70'in altında kaldığını göstermektedir. **Sonuç:** Bu makale, demografik veriler gibi KYTA tarafından sağlanan bilgilerin, bir bireyin gelecekte sigara içmesi olasılığının hesaplanmasına yardımcı olabileceğini ortaya koymaktadır.

Anahtar kelimeler: Sigara içmek, tütün kullanımı, kamu sağlığı

Introduction

The collection of data has become an easier process along with the rapid development of technology. A significant amount of data is available in science, industry, business, and many other areas in today's world. Tobacco use and control are also one of the most important research fields where enormous data has been collected recently. After the entrance of the World Health Organization Framework Convention on Tobacco Control (WHO FCTC) into force, many countries started to conduct Global Adult Tobacco Surveys (GATSs) and Global Youth Tobacco Surveys (GYTSs) regularly to monitor the prevalence of tobacco use and the effects of key tobacco control measures. The data supplied by these surveys builds considerable datasets for smoking issues. These datasets can be used for transforming the collected data to valuable information using data mining methods in order to help decision makers.

Data mining can be defined as the process of extracting knowledgeable information in an understandable structure from very large amounts of data.[1] It has become one of the most popular disciplines of applied science [2] due to its capability of discovering hidden patterns [3] in data. Classification is one of the important functions of data mining that classifies a data item into one of the different pre-defined classes.

This paper mainly conducted classification analyses on the GATSs data using three different classification algorithms to analyze a) how the GATSs data can be used for extracting valuable information to understand the relations between some important factors and smoking status of people and b) the prediction performance of the implemented classification algorithms on the GATS data.

In the literature, various data mining methods have been applied to various datasets for several different research fields. However, there are also few studies that focus on tobacco research area. For example; Montaño-Moreno et al.[4] used *multilayer perceptron*, *radial basis function*, *probabilistic neural network* and etc. to analyze the predictive power of different psychosocial and personality variables on the nicotine consumption of teenagers while Moon et al.[5] applied *decision tree models* to characterize smoking behavior among older adults considering the psychological distress, health status, alcohol use, and demographic variables. In the studies of Ding et al.[6] and Yun et al.[7], different algorithms such as *neural network*, *decision tree*, and etc. were used to examine quitting behaviors of people.

Some of the studies that used different datasets were also provided in this study. Sofean and Smith8 and Myslín et al.[9] smoking behaviors of people using data provided by Twitter while Benjakul et al.10 performed a clustering analysis to examine the characteristics of manufactured and roll your own cigarette users using data provided by the GATS 2009 Thailand. Nollen et al.[11] also explored the relations between demographic, psychosocial factors, and tobacco to determine cigarette smokers at higher risk for alternative tobacco product use from a diverse sample of adult smokers. In 2019, Singh and Katyan [12] analyzed the GATS 2010 data to characterize nicotine dependency using decision tree approach.

Apart from these studies, there are several different types of research which used data mining methods on smoking issues at the point of medical care. In 2015, Ding et al.[13] performed a classification analysis based on Support Vector Machine using structural Magnetic Resonance Imaging (MRI) images whereas McCormick et al.[14] classified the patient smoking status using semantic features of patients. In addition, Figueroa et al.[15] used clinical narrative texts to extract smoking status of patients while Wicentowski and Sydes[16] used implicit information from medical discharge summaries of patients. In the study of Sordo and Zeng[17], the dependency among sample size and classification performance of *Naive Bayes*, *Support Vector Machines*, and *Decision Trees* were examined using data supplied by patients. On the other hand, Huang et al.[18] examined the prescribing of smoking cessation medications in the primary care using rule mining methods.

In the light of the brief literature review provided above, it can be seen that there is a limited number of papers that studied the GATSs data using different kinds of data mining algorithms.

In this study, three well-known classification methods: "*K nearest neighbor (KNN)*", "*C4.5 algorithm*", and "*multilayer perceptron*" have been applied to the GATS Turkey 2012 data to classify the smoking status of the participants (two pre-defined classes: no smoker and daily smoker) based on some of their fundamental socio-demographic characteristics (age, gender, residence, education level, and working status). Additionally, the performance of the algorithm that provided the best outputs for Turkey case was tested using the data of six different European countries (Greece, Kazakhstan, Poland, Romania, Russia, and Ukraine) which locate in the same WHO region (Europe) with Turkey and provided open access to their GATSs data via the web page of Center for Disease Control and Prevention (CDC) during the study period.

The GATS is one of the most important surveys that provides vast body of data demographic characteristics, tobacco use behaviors and opinions for tobacco control policies of participants. It is also supported by WHO and implemented by several different countries over years. Many countries has been used this survey to monitor tobacco use and observe the performance of tobacco control policies. Therefore, this survey has become one of the most helpful surveys that researchers of tobacco field need. To our best knowledge, the data of these surveys mostly analyzed with survey methodologies. However, advanced methods can also help to reveal hidden knowledge that can increase our understanding on the relevant field. In this study, the relations between different demographic characteristics of people and their tobacco use behaviors were investigated with different data mining algorithms. This study is an important example how different data mining algorithms can be used on this survey.

The remainder of this paper was organized as follows. First, the methodology was discussed. Then, the results were presented. Finally, conclusions and discussions were provided.

Methods

This study has several steps as described in Figure 1. The initial classification analysis was conducted using the GATS data from Turkey (2012). The GATS is a national household survey[19] that helps nations to collect data on the prevalence of tobacco use and key tobacco measures.[20] It also covers data on some of the fundamental socio-demographic characteristics of the participants. In the content of this study, five *easy to reach* and *well known* demographic factors: *age*, *gender*, *residence*, *education level*, and *working status* were selected to perform analyses.

The GATS Turkey 2012 was performed with a total of 9851 participants. However, some participants did not respond to the selected demographic questions. A total of 24 participants did not declare the work status while 2 participants did not provide education level information. For that reason, these participants were excluded from the performed study and the data of 9825 participants were used for the further analyses. Before considering all of these candidate factors to be considered in our analyses; the dependency between the factors and the current smoking status of people were analyzed using *Chi Square Test*. Subsequently, the significantly depended factors were included for classification purposes.

The corresponding questions and responses used in GATS in 2012 are listed below. The frequency and percentages of the used data were also provided in Table 1.

- *Age: Respondents age in years?* The age data of the participants were collected as numeric variables. In this study, we categorized the ages of the people in 4 classes: 15-24, 25-44, 45-64, 65+. This classification was also used by WHO while analyzing the results of the GATSs.
- *Gender: Gender? (Male and female)*
- *Residence: Residence status? (urban and rural)*
- *Education levels: What is the highest education you have completed? (not graduated, elementary school, primary school, secondary or vocational school, high school, college or faculty, and master or Ph.D.)* Education levels of the countries were collected in 3 classes:no formal schooling (not graduated), primary education (primary to high school), higher education (university, MSc, and PhD in this study.
- *Working status: Which of the following best describes your main work status over the past 12 months? (paid employee, self-employed, non-paid family worker, student, homemaker, retired, no job (not able to work), and no job (able to work)).*Working status is collected in 5 classes: employee or employer, student, homemaker, retired and unemployed in this study.
- *Smoking status: Do you currently smoke tobacco on a daily basis, less than daily, or not at all? (daily, less than daily, and not at all).* In this study, the smoking statuses are defined in two classes: smoker (daily and less than daily smokers) and no smokers.

In this study, three different machine learning algorithms were used to perform a detailed classification analyses. During the selection of the types of algorithms, the main approaches that the algorithms have been used were investigated and algorithms which basically use different approaches from each other were selected for the further analyses. Therefore, *KNN, multilayer perceptron* and *C4.5 algorithms* were implemented using the software WEKA (Waikato Environment for Knowledge Analysis) which provides a collection of machine learning algorithms with single user interface.[21] These classification methods are known to be compatible with the GATS data. *KNN* algorithm performs a case base learning while *C4.5* constructs a decision tree and *multilayer perceptron* maps sets of input data onto a set of appropriate outputs. Brief information about these methods

Table 1. The statistics of the data supplied from GATS 2012 Turkey

Demographic Characteristics	Sub-categories	n	%
Age Group	*15-24*	1275	12.97
	25-44	3945	40.15
	45-64	2987	30.40
	65+	1618	16.46
Gender	*Male*	4453	45.32
	Female	5372	54.67
Residence	*Urban*	4912	49.99
	Rural	4913	50.00
Education Level	*Not Graduated*	1832	18.64
	Primary Education	6915	70.38
	Higher Education	1078	10.97
Work Status	*Employee or Employer*	3584	36.47
	Student	566	5.76
	Homemaker	3832	39.00
	Retired	1338	13.61
	Unemployed	505	5.13

was also provided in the next sub-sections. There are also some other reasons to chose these algorithms. These algorithms have been used on different datasets in several different areas and provided promising results. They are easy to understand, reach and implement.

In order to evaluate the classification performance of the algorithm giving the best classification result for Turkey case, the data sets of six different countries were also analyzed. The selected countries namely Greece, Kazakhstan, Poland, Romania, Russia, and Ukraine locate in the same WHO region (Europe) with Turkey and provide open access to their GATSs data from the web page of CDC. The data for these countries belongs to different years since GATS was performed in different years. Thus, GATS data from Greece belongs to the year 2013 with 4352 participants, from Kazakhstan belongs to the year 2014 with 4404 participants, from Poland belongs to the year 2009-2010 with 7786 participants, from Romania belongs to the year 2011 with 4488 participants, from Russia belongs to the year 2016 with 11440 participants, and from Ukraine belongs to the year 2017 with 8227 participants. This study was performed with the given years for the data. For that reason, the references were taken according to the year at which the analyses of the study was performed.

K-Nearest Neighbor (KNN)

K nearest neighbor (KNN) is one of the effective machine learning methods which is also known as instance-based learning, case-based learning, lazy learning.[1] The nearest algorithms are simply select the training instances with the closest distance to the query instance.[22] It has only one parameter which is called as k, number of neighbors.[23] Thus, as the nearest neighbor algorithm, KNN firstly; trains a set of cases and when a new case is needed to classify, it finds k number of training cases closest to the new point using a similarity function (such as Euclidean distance).[24] KNN can be advantageous when the study will be performed with the small database because the speed of computing distance will increase according to the number of instances.[22]

C4.5 Algorithm

The C4.5 algorithm was developed by Ross Quinlann, is a classification algorithm producing decision tree. It simply constructs a decision tree that is a predictive machine learning model[25] until it reaches the

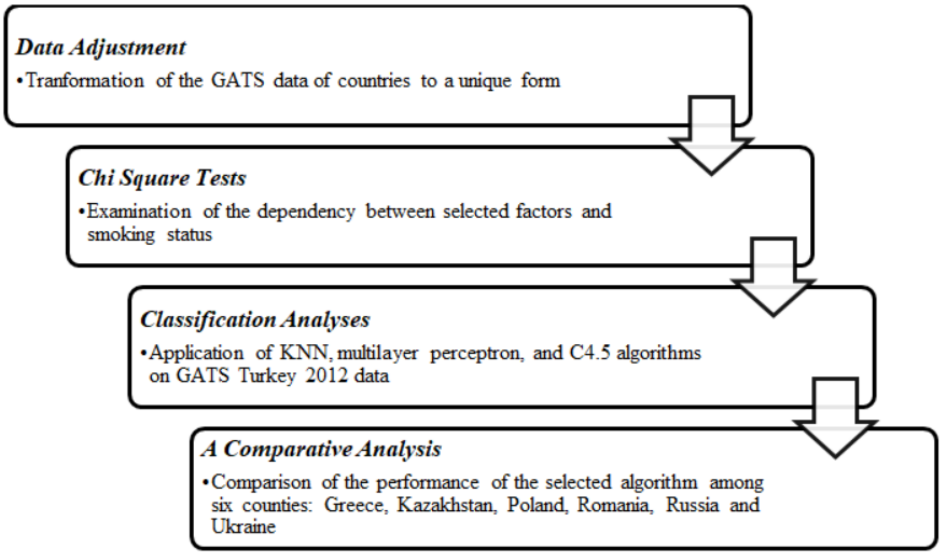
Data Adjustment
•Tranformation of the GATS data of countries to a unique form
Chi Square Tests
•Examination of the dependency between selected factors and smoking status
Classification Analyses
•Application of KNN, multilayer perceptron, and C4.5 algorithms on GATS Turkey 2012 data
A Comparative Analysis
•Comparison of the performance of the selected algorithm among six counties: Greece, Kazakhstan, Poland, Romania, Russia and Ukraine

Figure 1. The flow of the analyses performed in the presented paper.

equilibrium of flexibility and accuracy.[26] The internal nodes of the tree represent the different attributes while the branches between the nodes present the possible values.[25] Trees help researchers to determine useful predictors of an outcome efficiently and extract interactions between predictors without specifying these in advance.[27] The tree format of the algorithm allows generated rules to be easily interpreted and reduce the probability of errors.[28] They have provided useful results in medical field for disease diagnosis.[29]

C4.5 is known as a J48 algorithm in the Weka data mining tool. J48 is an open source Java implementation of the C4.5 algorithm in the Weka.30

Multilayer Perceptron

Multilayer Perceptron is one of the well-known neural network models[31] due to its clear architecture and comparably simple algorithm.[32] It is also a back-propagation algorithm[33] that conducts learning on a multilayer feed forward network.[34] Multilayer Perceptron consists of a number of neurons that are connected by weighted links.[35] In this algorithm, when data are denoted in the input layer, the network node conducts computations in the successive layers until an output value is reached at each of the output nodes.[32]

Results

The dependency among selected demographic factors and the smoking status of individuals were primarily tested using Chi Square Tests. For this aim 5 different hypotheses were prepared. An example for the hypothesis is given below.

Ho: residence and smoking status are independent
H1: residence and smoking status are dependent
As it can be seen at Table 2, all analyzed characteristics were found related to each other (<0.01) with the current smoking status of the individuals. Thus, all characteristics were included in the classification analyses.

Table 2. p values of the chi square tests.

	Residence	Age Group	Gender	Education Level	Working Status
Smoking Status	<0.001	<0.001	<0.001	<0.001	<0.001

For the classification analyses, two classes were pre-defined: smoker (class 1) and no smoker (class 2). All analyses were performed using a *10-fold cross validation* (k-fold cross validation) procedure that allow the effective use of the data.[24] In k-fold cross validation;

firstly, the data set is divided into k folds or subsets, secondly, one of the k folds is used as the test sets while k-1 subsets are used for training in turn, and finally, the average error for all k trials is calculated.[36]

The performances of the employed algorithms for Turkey case are compared by using percentages of the correctly classified instances, the values of the precision, the recall, F-measure for each class, and time is taken to build the model. The probability of *correct classification* is a performance measure that corresponds to the area under ROC curve.[37] *Precision* (that is also known as confidence) is the proportion of predicted positive instances that are correctly real positives while *recall* (that is also known as sensitivity) is the proportion of real positive instances that are correctly predicted positive.[38] The formulations of the recall and the precision are given in 1 and 2. On the other hand, *F-measure* can be defined as the harmonic mean of recall and precision.[39]

Precision (confidence)[40]= $\dfrac{\text{True Positive}}{\text{True positive+False Positive}}$

Recall (sensitivity)[40]= $\dfrac{\text{True Positive}}{\text{True positive+False Positive}}$

The results are as given at Table 3.

The findings of the classification analyses showed that C4.5 algorithm classified the instances correctly with the highest percentage (76.977%) for Turkish participants. Multilayer Perceptron had been the second best while KNN took the last place.

All of the tree algorithms were more confident and sensitive to classify no smoker class (class 2). The results showed that the performance measures for "no smoker" class for all algorithms were higher than 80%.

When the time taken to build model was searched, KNN algorithm took the first place with 0.00 seconds. C4.5 followed it with 0.08 seconds. Multilayer perceptron required more time to build model compared the other two algorithms. Thus, if the rate of correctly classified instances was important for decision makers, C4.5 algorithm dominated the other algorithms with the high correctly classification rate. On the other hand, if the speed of the analysis was primary for them, KNN algorithm was the best among three methods. In this study, the correct classification rate was the vital factor for us. Correspondingly, the C4.5 algorithm had the first place for our analyses. A detailed decision tree was obtained as an output of the C4.5 algorithm. Although the entire decision tree was too big to represent in one figure, only one section of the tree was represented in this paper (as seen in figure 2).

Table 3. Performance measures of three different classification methods on Turkey case.

		KNN	C4.5	Multilayer Perceptron
Correctly classified instances (%)		76.743	76.977	76.875
Precision	Class 1 (smoker)	0.536	0.548	0.534
	Class 2 (no smoker)	0.816	0.811	0.828
	Weighted Average	0.748	0.746	0.756
Recall	Class 1 (smoker)	0.381	0.348	0.439
	Class 2 (no smoker)	0.893	0.907	0.876
	Weighted Average	0.767	0.770	0.769
F value	Class 1 (smoker)	0.446	0.426	0.482
	Class 2 (no smoker)	0.853	0.856	0.851
	Weighted Average	0.753	0.751	0.761
Time is taken to build model (sec)		0.00	0.08	12.53
Total Num. of Instances	9825	9825	9825	9825

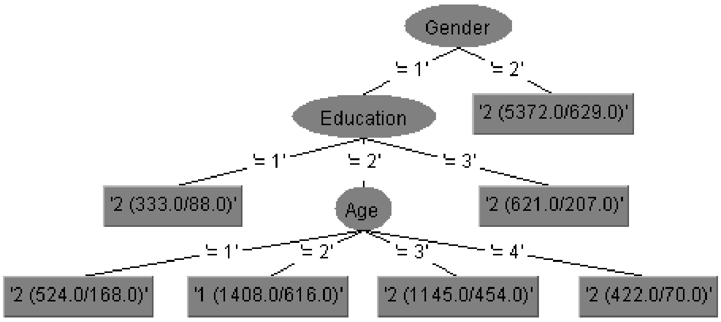

Figure 2. The C4.5 decision tree.

The algorithm started to classify the smoking status of people according to their genders. Males (represented with "1" in the figure) divided by new branches while all females (represented with "2" in the figure) were categorized directly as no smokers. Thus, we may say that this decision tree is more capable of classifying smoking status of males compared to females.

The algorithm continued to classify the smoking status of males according to their education levels. Males who are not graduated from any school (represented by 1) or graduated from higher education (BSc, MSc, PhD) (represented by 3) were classified as no smoker by the algorithm. Males who took primary education (represented by 2) were categorized according to their age group. The algorithm classified the males (who are primarily educated) aged among 15-24 (represented by 1), 45-64 (represented by 3), and 65+ (represented by 4) as no smoker while males aged 25-44 as smoker.

Finally, the findings belonging to Turkey were compared with six other European countries by implementing the best performing algorithm (C4.5). The C4.5 algorithm implemented for the GATS data from Greece 2013, Kazakhstan 2014, Poland 2009-2010, Romania 2011, Russia 2016, and Ukraine 2017. Table 4 exhibits the results of classification analysis of C4.5 algorithm for these countries. The findings indicated that C4.5 produce the highest correctly classification rate for Ukraine (80.369%). Kazakhstan followed Ukraine with 78.133%. Among countries, Greece had the lowest (68.910 %). Turkey took the fourth place after Romania with 76.977%.

Table 4. Performance measures of C4.5 algorithm for different countries.

		Greece	Kazakhstan	Poland	Romania	Russia	Ukraine
Correctly classified instances (%)		68.910	78.133	69.676	77.473	75.542	80.369
Precision	Class 1 (smoker)	0.616	0.531	0.584	0.583	0.598	0.569
	Class 2 (no smoker)	0.722	0.840	0.701	0.785	0.805	0.840
	Weighted Average	0.681	0.768	0.665	0.738	0.744	0.782
Recall	Class 1 (smoker)	0.498	0.435	0.064	0.127	0.489	0.357
	Class 2 (no smoker)	0.808	0.885	0.980	0.972	0.865	0.926
	Weighted Average	0.689	0.781	0.697	0.775	0.755	0.804
F value	Class 1 (smoker)	0.551	0.478	0.115	0.208	0.538	0.439
	Class 2 (no smoker)	0.762	0.862	0.817	0.869	0.834	0.881
	Weighted Average	0.681	0.773	0.600	0.714	0.748	0.786
Total Num. of Instances		4352	4404	7786	4488	11440	8227

Even though %76.977 correctly classification rate was not so high enough to make certain judgments about the smoking status of people, we should remind that the main purpose of this study was to classify the smoking status only with the limited number of socio-demographic characteristics of the individuals. Hence the obtained results can be acceptable considering the limits determined in the presented study.

Discussion

Tobacco use is still a prevalent issue[41] that treats the world population. In order to understand the behavior of smokers, countries collects considerable amount of data with the help of the WHO. This study focused on how the GATSs data can be used for extracting valuable knowledge about smoking related facts. Our main concern has been seeking for a relation between the smoking status and socio-economic factors. Initially, three different classification algorithms: KNN, multilayer perceptron, and C48 algorithms were used on the GATS data from Turkey (2012). Subsequently, the algorithm that provided the best classification results for Turkey was also used for other six European countries: Greece, Kazakhstan, Poland, Romania, Russia, and Ukraine to evaluate the performance of the algorithm on different data sets.

The outputs of the analyses indicated that C4.5 algorithm classified the instances of Turkey more correctly than other two algorithms. That is why; the C4.5 algorithm was used for the classification of the smoking status of individuals for Greece, Kazakhstan, Poland, Romania, Russia, and Ukraine. Ukraine had the highest correctly classification rate among them while Greece had the lowest. The results mainly showed that Ukraine, Kazakhstan, Romania, and Turkey had considerable classification performance for the C4.5 algorithm when compared to others.

The findings of the analyses indicated that the smoking status of approximately 80% of GATS participants was correctly classified by using socio-demographic factors. The best performing algorithm (C4.5) for Turkey was found to be much more capable of classifying the smoking status of males. One of the main reasons of this fact can be the lower number of female smokers in the studied sample. The algorithm could classify the male participants according to education level and age group. Thus, some characteristics such as education level and age group may be accepted as more influential factors compared to others. This may show us that the data about socio-demographic characteristics provided by GATSs can be used as a clue for prediction of smoking status of individuals by decision makers. Thus, this paper showed a convenient application how the GATS data can be used for different purposes besides monitoring the prevalence of tobacco use and the effects of key tobacco control measures. The findings of the study can be helpful to understand the relationships between the smoking status of the individuals and their fundamental characteristics. The findings can also be used to compute the likelihood of an individual to be a smoker in the future. Thereby, some of the smoking cessation policies can be adjusted according to the different age and education groups. Executing the different policies for different groups is expected to be more effective (less cost and time) when compared to implementation of a general set policy.

Hence, conducting detailed analyses with advanced data mining methods using the GATSs data can increase knowledge on smoking issues. Conversion the GATSs data to a more valuable and understandable structure may be beneficial for decision makers and policy makers to use the new form of the data in order to provide scientific evidence for future decision support. However, the performances of the algorithms can change according to the studied database. For example; C4.5 algorithm classified the instances in Turkey case better than Greece case. That is why; it is important to keep in mind that testing the performance of the different algorithms is crucial to extract valuable knowledge from the GATSs data.

This study has also some limitations. In the content of this study, the classification performances of the three data mining methods were tested. Different classification algorithms can also provide better or worse results than the performed analyses for the studied cases. The algorithm which had the highest correctly classification rate for this study was only capable of classifying the smoking status of males in detail. Studying with different algorithms may also provide a comprehensive classification for females, too and overcome this problem of the performed study. Moreover, the GATS Turkey data was used for the analyses. The comparison analysis was conducted only with countries located in WHO Europe region. The performance of the algorithms can also be tested with the data of other world countries to obtain a vast frame for this topic.

Acknowledgment

Financial Support: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of Interest: The authors have no conflicts of interest.

Contributorship Statement: Both authors participated in the design, analysis and interpretation of the study. Pınar Kocabey Ciftci provided the research idea and adjusted the required data. Zeynep Didem Unutmaz Durmusoglu searched and chose the implemented data mining algorithms. Both authors performed the analysis and interpreted the results of the study.

References

1. Jabbar MA, Deekshatulu BL, Chandra P. Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm. Procedia Technol. 2013;10:85-94.
2. Kartelj A. Classification of Smoking Cessation Status Using Various Data Mining Methods. Math Balk New Ser. 2010;24(3-4):199-205.
3. Segall RS, Guha GS, Nonis SA. Data mining of environmental stress tolerances on plants. Kybernetes. 2013;37:127-148.
4. Montaño-Moreno JJ, Gervilla-García E, Cajal-Blasco B, Palmer A. Data mining classification techniques: an application to tobacco consumption in teenagers. An Psicol. 2014;30(2):633-641.
5. Moon SS, Kang S-Y, Jitpitaklert W, Kim SB. Decision tree models for characterizing smoking patterns of older adults. Expert Syst Appl. 2012;39(1):445-451.
6. Ding X, Bedingfield S, Yeh C-H, et al. Identifying Tobacco Control Policy Drivers: A Neural Network Approach. In: Leung CS, Lee M, Chan JH, eds. Neural Information Processing. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2009:770-776.
7. Yun C-J, Ding X, Bedingfield S, et al. Performance Evaluation of Intelligent Prediction Models on Smokers' Quitting Behaviour. In: Fyfe C, Kim D, Lee S-Y, Yin H, eds. Intelligent Data Engineering and Automated Learning – IDEAL 2008. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2008:210-216.
8. Sofean M, Smith M. Sentiment analysis on smoking in social networks. Stud Health Technol Inform. 2013;192:1118.
9. Myslín M, Zhu S-H, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res. 2013;15(8):e174.
10. Benjakul S, Termsirikulchai L, Hsia J, et al. Current manufactured cigarette smoking and roll-your-own cigarette smoking in Thailand: findings from the 2009 Global Adult Tobacco Survey. BMC Public Health. 2013;13:277.
11. Nollen NL, Ahluwalia JS, Lei Y, Yu Q, Scheuermann TS, Mayo MS. Adult Cigarette Smokers at Highest Risk for Concurrent Alternative Tobacco Product Use Among a Racially/Ethnically and Socioeconomically Diverse Sample. Nicotine Tob Res Off J Soc Res Nicotine Tob. 2016;18(4):386-394.
12. Singh A, Katyan H. Classification of nicotine-dependent users in India: a decision-tree approach. *J Public Health*. 2019;27(4):453-459.
13. Ding X, Yang Y, Stein EA, Ross TJ. Multivariate classification of smokers and nonsmokers using SVM-RFE on structural MRI images. Hum Brain Mapp. 2015;36(12):4869-4879.
14. McCormick PJ, Elhadad N, Stetson PD. Use of semantic features to classify patient smoking status. AMIA Annu Symp Proc. 2008;2008:450-454.

15. Figueroa RL, Soto DA, Pino EJ. Identifying and extracting patient smoking status information from clinical narrative texts in Spanish. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. ; 2014:2710-2713.
16. Wicentowski R, Sydes MR. Using Implicit Information to Identify Smoking Status in Smoke-blind Medical Discharge Summaries. J Am Med Inform Assoc JAMIA. 2008;15(1):29-31.
17. Sordo M, Zeng Q. On Sample Size and Classification Accuracy: A Performance Comparison. In: Oliveira JL, Maojo V, Martín-Sánchez F, Pereira AS, eds. Biological and Medical Data Analysis. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2005:193-201.
18. Huang Y, Britton J, Hubbard R, Lewis S. Who receives prescriptions for smoking cessation medications? An association rule mining analysis using a large primary care database. Tob Control. 2013;22(4):274-279.
19. Kaleta D, Usidame B, Dziankowska-Zaborszczyk E, Makowiec-Dąbrowska T. Socioeconomic Disparities in Age of Initiation and Ever Tobacco Smoking: Findings from Romania. Cent Eur J Public Health. 2015;23(4):299-305.
20. WHO [Internet]. Global Adult Tobacco Survey (GATS). WHO. [Cited:14.06.2016]. Avaliable from: http://www.who.int/tobacco/surveillance/survey/gats/en/
21. Hussain S, Alili AA. A pruning approach to optimize synaptic connections and select relevant input parameters for neural network modelling of solar radiation. Appl Soft Comput. 2017;52:898-908.
22. Tou JY, Tay YH, Lau PY. Recent trends in texture classification: A review. In: Symposium on Programs in Information & Communication Technology. 2009; 2(3).
23. Liao TW, Kuo RJ. Five discrete symbiotic organisms search algorithms for simultaneous optimization of feature subset and neighborhood size of KNN classification models. Appl Soft Comput. 2018;64:581-595.
24. Amaral JLM, Lopes AJ, Jansen JM, Faria ACD, Melo PL. An improved method of early diagnosis of smoking-induced respiratory changes using machine learning algorithms. Comput Methods Programs Biomed. 2013;112(3):441-454.
25. Classification Methods [Internet]. [Cited: 13.06.2017]. Avaliable from: http://www.d.umn.edu/~padhy005/Chapter5.html
26. Kaur G, Chhabra A. Improved J48 Classification Algorithm for the Prediction of Diabetes. Int J Comput Appl. 2014;98(22):13-17.
27. King MW, Resick PA. Data mining in psychological treatment research: a primer on classification and regression trees. J Consult Clin Psychol. 2014;82(5):895-905
28. Hsu-Che W, Ya-Han H, Yen-Hao H. Two-stage credit rating prediction using machine learning techniques. Kybernetes. 2014;43(7):1098-1113.
29. Zhu Y, Fang J. Logistic Regression–Based Trichotomous Classification Tree and Its Application in Medical Diagnosis. Med Decis Making. 2016;36(8):973-989.
30. Gholap J. Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility. Asian J Comput Sci Inf Technol. 2012;2(8). Accessed August 19, 2016. http://arxiv.org/abs/1208.3943
31. Nadeem M, Banka H, Venugopal R. Estimation of pellet size and strength of limestone and manganese concentrate using soft computing techniques. Appl Soft Comput. 2017;59:500-511.
32. Yan H, Jiang Y, Zheng J, Peng C, Li Q. A multilayer perceptron-based medical decision support system for heart disease diagnosis. Expert Syst Appl. 2006;30(2):272-281.
33. Malhotra R. An empirical framework for defect prediction using machine learning techniques with Android software. Appl Soft Comput. 2016;49:1034-1050.
34. Arora R, Suman. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. Int J Comput Appl. 2012;54:13.

35. Riedmiller M. Advanced supervised learning in multi-layer perceptrons — From backpropagation to adaptive learning algorithms. Comput Stand Interfaces. 1994;16(3):265-278.
36. Cross Validation [Internet]. [Cited:19.08.2016]. Avaliable from: https://www.cs.cmu.edu/~schneide/tut5/node42.html
37. Steinbach WR, Richter K. Multiple Classification and Receiver Operating Characteristic (ROC) Analysis. Med Decis Making. 1987;7(4):234-237.
38. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation - SIE07001.pdf. Accessed August 19, 2016. https://csem.flinders.edu.au/research/techreps/SIE07001.pdf
39. Danenas P, Garsva G. Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach. Procedia Comput Sci. 2012;Complete(9):1324-1333.
40. WekaDataAnalysis [Internet]. [Cited: 15.07.2016]. Avaliable from: http://www.cs.usfca.edu/~pfrancislyon/courses/640fall2015/WekaDataAnalysis.pdf
41. Mermer G, Dağhan Ş, Bilge A, Dönmez RÖ, Özsoy S, Günay T. Prevalence of Tobacco Use among School Teachers and Effect of Training on Tobacco Use in Western Turkey. Cent Eur J Public Health. 2016;24(2):137-143.