BMIJ
ISSN: 2148-2586

bmij (2023) 11 (3): 1152-1162

doi: https://doi.org/10.15295/bmij.v11i3.2293

Research Article

A hybrid machine learning approach in predicting smoking behaviour: The case of Turkey

Sigara içme davranışını tahmin etmede hibrit bir makine öğrenimi yaklaşımı: Türkiye örneği

Nuray Tezcan[1]

Gökçe Karahan Adalı[2]

Anıl Burcu Özyurt Serim[3]

[1] Prof. Dr., Haliç University, Istanbul, Turkey, nuraytezcan@halic.edu.tr

ORCID: 0000-0002-3184-7330

[2] Assist. Prof., Haliç University, Istanbul, Turkey, gokceadali@halic.edu.tr

ORCID: 0000-0001-8567-4626

[3] Assist. Prof., Haliç University, Istanbul, Turkey, burcuozyurt@halic.edu.tr

ORCID: 0000-0001-9868-2676

Corresponding Author:

Gökçe Karahan Adalı,

Haliç University, İstanbul, Turkey
gokceadali@halic.edu.tr

Submitted: 11/08/2023

Revised: 17/09/2023

Accepted: 20/09/2023

Online Published: 25/09/2023

Citation: Tezcan, N., & Karahan Adalı, G., & Özyurt Serim, A.B., A hybrid machine learning approach in predicting smoking behaviour: The case of Turkey, bmij (2023) 11 (3): 1152-1162, doi: https://doi.org/10.15295/bmij.v11i3.2293

Abstract

This study aims to analyze the smoking behaviour of people aged 15 and older in Turkey using supervised and unsupervised machine learning methods. In this study, C4.5 and Random Forest (RF) were trained to predict smoking behaviour, and an apriori algorithm was used to detect associations. Sensitivity, specificity, accuracy, positive predicted value (PPV), and f-measure were used to compare the performances of the supervised models. The Turkey Health Interview Survey 2019 was used with a sample size of 17084 to predict smoking behaviour and determine the factors affecting smoking. Data analysis and performance evaluation were performed with R programming language by RStudio. By association rules, gender, age, and alcohol consumption are the most representative attributes of smoking behaviour. Associations were determined on smoking, non-smoking and quit-smoking behaviour. Also, it has been seen that the RF algorithm has better results than the C4.5 algorithm. It's preferred to use the RF model, which had better performance with an accuracy of 0.909, a specificity of 0.965, a sensitivity of 0.782, a PPV of 0.908, and an f-measure of 0.840 for predicting smoking behaviour. This study contributes to the literature covering the most comprehensive national health survey data and using machine learning methods on this data in Turkey. Also, it indicates that machine learning methods can be used to analyze such datasets.

Keywords: Health Behaviour, Cross-Sectional Models, General

Jel Codes: I12, C21, C8

Öz

Bu çalışma, Türkiye'de 15 yaş ve üstü kişilerin sigara içme davranışlarını hem denetimli hem de denetimsiz makine öğrenimi yöntemleri kullanarak analiz etmeyi amaçlamaktadır. Bu çalışmada, sigara içme davranışını tahmin etmek için C4.5 ve Random Forest (RF) eğitilmiş ve ayrıca ilişkileri tespit etmek için apriori algoritması kullanılmıştır. Denetlenen modellerin performanslarını karşılaştırmak için duyarlılık, özgüllük, doğruluk, pozitif tahmin değeri (ppv), f-ölçüsü kullanıldı. Sigara içme davranışını tahmin etmek ve sigarayı etkileyen faktörleri belirlemek için 17084 örneklem büyüklüğü ile Türkiye Sağlık Araştırması 2019 kullanılmıştır. Veri analizi ve performans değerlendirmesi RStudio tarafından R programlama dili ile yapılmıştır. Birliktelik kurallarına göre cinsiyet, yaş ve alkol tüketimi, sigara içme davranışının en temsili özellikleri olarak belirlenir. Dernekler sigara içme, içmeme ve sigarayı bırakma davranışında belirlendi. Ayrıca RF algoritmasının C4.5 algoritmasına göre daha iyi sonuçlara sahip olduğu görülmüştür. Sigara içme davranışını tahmin etmede 0,909 doğruluk, 0,965 özgüllük, 0,782 duyarlılık, 0,908 ppv, 0,840 f-ölçüsü ile daha iyi performans gösteren RF modelinin kullanılması tercih edilmiştir. Bu çalışma hem Türkiye'deki en kapsamlı ulusal sağlık araştırması verilerini kapsaması hem de bu veriler üzerinde makine öğrenmesi yöntemlerinin kullanılması açısından literatüre katkı sağlamakta ve bu tür veri setlerinin makine öğrenimi yöntemleriyle analiz edilebileceğini göstermektedir.

Anahtar Kelimeler: Davranışsal Sağlık, Yatay Kesit Modelleri, Genel

JEL Kodları: I12, C21, C8

Introduction

Tobacco use is one of the most persistent and troublesome public health concerns, despite the efforts of nations to develop programs and regulations that result in remedies (World Health Organization-WHO, 2021). Reducing tobacco use is an effective way to prevent noncommunicable diseases (NCDs), which cause 71% of all fatalities worldwide (WHO, 2021; WHO, 2022). For that reason, it is one of the targets in the Sustainable Development Goals (SDGs) framework that has been adopted throughout the world (United Nations, 2015)

In addition to being important for global health, reducing tobacco use is also important for sustainable economic growth. It is estimated that the annual cost of smoking drains the world economy of about US$1.4 trillion (Goodchild, Nargis and d'Espaignet, 2018). High tobacco use, however, endangers sustainable development by making impoverished households even poorer due to rising healthcare expenditures and declining revenues. In 2003, WHO Member States ratified a convention known as the WHO Framework Convention on Tobacco Control (WHO FCTC). The Convention provides a framework for parties to implement tobacco control measures to continuously and significantly lower the prevalence of tobacco use and exposure to tobacco smoke.

The goal of the road map outlined in the WHO Global Action Plan 2013-2020 is to reduce deaths from NCDs by 25% by the year 2025. Current trends suggest that by 2030, it will cause more than 8 million deaths annually, according to the World Health Organization's 2020 report (WHO, 2021). Within the scope of that report, the male-female ratio of current tobacco use among people aged 15 and over varies by country, but this ratio is on the decline. The greatest rates of tobacco use are found among men between the ages of 45 and 54, according to the WHO's global study on trends in the prevalence of tobacco use between 2000 and 2025. Since women's tobacco consumption varies from year to year, on average, it peaks in the 55–64 age range. The majority of EU member states have been found to have the highest daily smoking rates between the ages of 25 and 54, with the lowest rates occurring after age 65. The lowest rates for men and women have always been 75 years of age and older in all EU member states, as well as Serbia and Turkey. It is predicted that tobacco use rates for both men and women will tend to decline until 2025 based on the age variable. (WHO, 2019)

This study examines the smoking behaviour of the respondents using a machine learning method based on the most comprehensive national health survey in Turkey. According to this, the next part of the study focuses on a literature review, and the third part provides information about material and methods. After providing the results of the analyses in the following section, discussions and conclusions are presented.

Literature review

Machine learning makes it easier to identify complex patterns in data sets for clinical research. Additionally, machine learning algorithms can be used to discover nonlinear relationships and novel features in data. Over the past two decades, several studies have been presented regarding this issue. Some of these are given in the following:

Classification Trees (CTs) are useful machine-learning algorithms that can be used to classify data on tobacco use, according to the majority of research (Abo-Tabik, Costen, Darby and Benn, 2019; Coughlin, Tegge ,Sheffer, Bickel, 2020; Koslovsky, Swartz,Chan,Leon-Novelo,Wilkinson,Kendzor and Businelle, 2018; Zhang, Liu, Zhang, Huang, 2019). Dumortier, Beckjord, Shiffman and Sejdić, 2016 evaluated the desire to smoke based on 41 characteristics using data gathered from university students. (e.g., alcohol consumption, mood status, hunger, location, type of work, etc.). This research compared the performance of three machine learning algorithms: Naive Bayes, Discriminant Analysis, and CTs. The findings showed that machine learning was highly accurate at forecasting smokers' propensity, with CTs outperforming other methods with an average performance of about 86%.

While McCormick, Elhadad, and Stetson (2008) assessed the smoking status of the patient using the semantic characteristics of the patients, Ding, Yang, Stein, and Ross (2017) performed a classification study based on the Support Vector Machine (SVM) utilizing structural Magnetic Resonance Imaging (MRI) data. Nollen, Ahluwalia, Lei, Yu, Scheuermann, and Mayo (2016) discovered that adults who smoke are more inclined to utilize novel alternative tobacco products. This study also examined whether there were any relationships between tobacco use and smokers' psychosocial and demographic traits.

Koslovsky, Swartz, Chan, Leon-Novelo, Wilkinson, Kendzor and Businelle (2018) used machine learning methods to comprehend smokers' behaviour and cravings that alter throughout the smoking cessation process. This method determined which variables were effective between quitting smoking and starting again. Singh and Katyan (2019) used a decision tree approach to characterize nicotine dependence using demographic and socio-economic variables. According to the results, duration of smoking, education, gender and region were important variables for smoking dependency. In contrast, duration of smokeless tobacco use, education, occupation and age were important variables for smokeless tobacco use.

Research on individuals' addictions by Mak, Lee and Park (2019) revealed that supervised learning methods such as CTs, Naive Bayes, logistic regression, SVM, and neural networks are more often utilized and outperform other unsupervised methods. According to a study by Maginnity (2020), the prediction of whether a person has used tobacco products in the last 30 days was evaluated between the logistic regression and RF classification models. This study aims to determine whether an RF model will have a higher prediction accuracy than its logistic regression alternatives. In conclusion, the logistic models slightly outperformed their RF counterparts, indicating that these classification models effectively identify adolescents who won't start using tobacco.

Durmuşoğlu and Kocabey (2021) classified the participants' smoking status as either non-smokers or daily smokers using the Global Adult Tobacco Survey (GATS) Turkey 2012 data set. The k-nearest neighbour (k-NN), C4.5 algorithm, and multilayer perceptron methods were used to find the best classification performance. This research used various data mining algorithms to explore the relationships between people's tobacco use behaviours and various demographic traits, including age, gender, place of residence, educational attainment, and employment status. As a result, the C4.5 decision tree method was discovered to be the best-performing algorithm, and it was observed that male participants had much better success classifying smoking status by education level and age group.

Abo-Tabik, Benn, and Costen (2021) used data gathered from cell phone sensors to model the smoking behaviour of individuals using three distinct supervised machine learning models CTs, SVM, and convolutional Neural Networks). When predicting a smoking event, this model demonstrated that the convolutional Neural Network (CNN) method worked better than other methods, with an accuracy of about 86.6%.

Thakur, Poddar, and Roy (2022) propose a machine learning-based modelling framework using sensor data to describe smoking activities in real-time. This research aimed to compare various classification algorithms and identify the most effective classification method to address this issue. The comparison of the various classification models' prediction capabilities makes it simple to choose the best classification model for the application. In order to classify data, methods like logistic regression (LR), k-NN, adaptive reinforcement, RF, SVM, and decision tree (DT) are employed.

According to a study by Jitenkumar Singh, Jiran Meitei, Alee, Kriina, and Haobijam (2022), the RF algorithm was superior and performed much better in predicting the status of smokeless tobacco use in women from northeastern Indian states than the other ML algorithms. It is thought that the research will make an important contribution to the literature as one of the examples showing how machine learning algorithms can be applied in the classification of health-related events.

Evenhuis, Occhipinti, Jones, and Wishart (2023) studied factors associated with smoking cessation in health professionals and found evidence that age and work environment factors predict suicide attempt success in some health professional groups.

Issabakhs ,Sánchez-Romero, Le, Liber, and Tan (2023) analysis revealed that more past 30 days of e-cigarette use at the time of quitting, fewer past 30 days of cigarette use before quitting, ages older than 18 at smoking initiation, fewer years of smoking, poly tobacco past 30 days use before quitting, and higher BMI resulted in higher chances of cigarette cessation for adult smokers in the US.

Material and methods

In this study, data analysis was performed using R programming language and RStudio (RStudio, 2019) as development tools for R codes[4] . In this study, C4.5 and Random Forest (RF) were trained to predict smoking behaviour and apriori algorithm was used to detect associations of individual and smoking behaviour characteristics. These models are explained in the modelling section deeply. In addition, the steps of the CRISP-DM (Cross-Industry Standard Process for Data Mining) model were followed across data analysis (Shearer, 2000). This model comprises six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

[4] The following R packages were used in the study: arules (Hahsler, 2023), arulesViz (Hahsler, 2023), caret (Kuhn, 2018), dplyr (Wickham, 2023), plyr (Wickham, 2022), randomForest (Liaw, 2022), RWeka (Hornik, Buchta, & Zeileis, 2009),), rJava (Urbanek, 2021), tidyVerse (Wickham, 2023).

Defining Problem: The problem was stated as applying machine learning techniques to predict smoking behaviour and identify attributes influencing smoking behaviour among Turkish adults aged 15 and older.

Data Understanding: The data set used in this research came from the Turkish Statistical Institute's (TSI) 2019 Turkey Health Interview Survey (TurkStat, 2019). The data collection phase was conducted by TSI in 2019, therefore ethical committee approval was not obtained by the authors.

This survey was the latest and most comprehensive national health survey when the study was employed. In this data set, there have been more than 200 attributes for all ages, and this survey included 23708 people; however, the individuals under 15 were dropped from the data set because this study examined people aged 15 years and over. Thus, the final data set comprised 17084 people. Except for smoking use behaviour, ten available questions for the study regarding the respondents' socio-economic and health status were selected as attributes. These attributes used in the analyses and their descriptions, frequencies and percentages of the categories are presented in Table 1.

Table 1: Attributes Used and Their Descriptions, Frequencies, and Percentages

Attribute	Type of Attribute	Description		Frequency	Percentage
Smoking	Nominal	1	Tobacco smokers	5231	30.62
		2	Non-smokers	9256	54.18
		3	Former tobacco-smokers	2597	15.20
Gender	Nominal	1	Male	7784	45.56
		2	Female	9300	54.44
Marital Status	Nominal	1	Single	3610	21.13
		2	Married	11726	68.64
		3	Divorced	574	3.36
		4	Widowed	1174	6.87
Age (years)	Ordinal	1	15-24	2730	15.98
		2	25-34	3070	17.97
		3	35-44	3395	19.87
		4	45-54	2918	17.08
		5	55-64	2513	14.71
		6	65-74	1590	9.31
		7	75+	868	5.08
Education	Ordinal	1	Non-literate	1371	8.03
		2	Primary school	6435	37.67
		3	Secondary school	2965	17.36
		4	High school	3246	19.00
		5	Undergraduate/Graduate	3067	17.95
Working Status	Nominal	1	Employee, Employer / self-employed	6126	35.86
		2	Working as unpaid family worker/Busy with the care of housework and/or family, child, elderly, sick, etc.	5965	34.92
		3	Job seeker	1003	5.87
		4	Continuing education/training	1303	7.63
		5	Retired/Leaving work life due to age-related reasons/Inoperable due to disabled and/or permanent health problems	2687	15.73
Income (Turkish Liras)	Ordinal	1	0-1668	2726	15.96
		2	1669-2424	3857	22.58
		3	2425-3398	3191	18.68
		4	3399-5052	4060	23.76
		5	5053-+	3250	19.02
General Health	Ordinal	1	Very good	1247	7.30
		2	Good	8741	51.16
		3	Fair	5214	30.52
		4	Bad	1655	9.69
		5	Very bad	227	1.33
Body Mass Index	Ordinal	1	Underweight	587	3.44
		2	Normal range	6581	38.52
		3	Overweight	6105	35.73
		4	Obese	3811	22.31
Daily Activity	Ordinal	1	Mostly sitting or standing	10873	63.64
		2	Mostly walking or tasks of moderate physical effort	5525	32.34
		3	Mostly heavy labour or physically demanding work	686	4.02
Alcohol Use	Nominal	1	Yes	4499	26.33
		2	No	12585	73.67

At first glance, it is clear that 54.4% of the participants are female and 45.6% are male. The ages of the participants are virtually evenly spread, and approximately 70% are married. The respondents' average level of education is not very high; approximately 55 % of them have completed both primary and secondary education. Additionally, 35% of the participants identified as employees or workers, while 35% were unemployed. Categories in terms of income level appear to have similar participants.

Data Preparing: In the Turkey Health Interview Survey 2019, the smoking status of the respondents is divided into four groups: Current daily smokers, current occasional smokers, non-smokers, and former smokers. However, the first and second groups were combined since the number of respondents in current occasional tobacco smokers is insufficient to reveal meaningful results. In addition to socio-economic attributes like gender, age, education, marital status, income, and employment status, this study also looked at general health, body mass index (BMI), daily activity status, and alcohol use to identify factors influencing smoking behaviour. Participants were questioned regarding their use of alcohol (even a little). Nominal or ordinal scales were used to measure all attributes.

Besides, categories of some attributes were rearranged to obtain appropriate categories for the analyses in this stage. For example, the number of categories was reduced from 11 to 5 for the education attribute, from 10 to 5 for the working status attribute, and from 20 to 5 for the income level attribute. With this approach, it was ensured that the categories contained more comprehensive information. As the BMI attribute was not included in the survey directly, it was calculated using respondents' weight and height values. BMI can be found by weight in kilograms divided by the square of height in meters for adults. Accordingly, four categories that are underweight, normal range, overweight and obese were obtained (WHO, 2023). The dataset has no missing value or outlier because all attributes used in the analyses are categorical data.

Modelling: This study's research problem was considered an association and prediction problem. For this purpose, association rule, C4.5 and RF methods were used, respectively.

The association rule method employed in the first part of the analysis process defines the probabilistic correlation between events. This correlation is obtained for the events that often occur together. Accordingly, the Apriori Algorithm for associations was applied to examine all three smoking behaviours; smokers, non-smokers, and quit-smokers were analyzed separately.

Apriori, a fundamental algorithm developed by R. Agrawal and R. Srikant in 1994 for mining frequent item sets for Boolean association rules, is the association approach that was chosen for this investigation. Because the method uses previous knowledge of common itemset qualities, the algorithm's name is based on this fact. A level-wise search is an iterative strategy Apriori uses to investigate (k+1)-itemsets using k-itemsets. By searching the database, adding up each item's count, and gathering the items that meet the minimal support, the frequent 1-item sets are first discovered. The set that results is known as L1. When no more frequent k-itemsets exist, L1 is utilized to locate L2, the collection of frequent 2-itemsets, and so on (Han & Kamber, 2006).

Finding a threshold value to determine the association rule is crucial to this concept. The researcher determines the threshold for support and confidence values; all associations are ranked according to this threshold. Additionally, some criteria should be established to differentiate between interesting and uninteresting association rules.

An objective measure for association rules of the form $X \Rightarrow Y$ is support, representing how frequently the items appear in the data. Support is the probability $P(XUY)$, where XUY denotes the union of the itemsets X and Y or the presence of both in a transaction.

Another measure is confidence, which refers to the percentage of transactions that contain a particular item or set of items. This is taken to be the conditional probability $P(Y|X)$, that is, the probability that a transaction containing X also contains Y. A third metric, called lift, is a simple correlation measure that is given as follows: A is independent of the occurrence of itemset B if $P(A \cup B) = P(A)P(B)$; In other words, it assesses the degree to which the occurrence of one “lifts” the occurrence of the other (Han&Kamber, 2016).

C4.5 is a decision tree formation technique, one of the most used algorithms for supervised machine learning classification. It achieves outstanding predicted performance (Han & Kamber, 2006). C4.5 can be regarded as an improved version of the ID3 algorithm, and it uses information gain known as gain ratio.

RF method is used as the second supervised machine algorithm in this study. It is a popular and very efficient algorithm formed by combining more than one decision tree and can be described as a tree-based ensemble learning approach. Therefore, the RF algorithm outperforms decision trees and solves classification and regression problems (Genuer & Poggi, 2020).

Evaluation: In this stage, the performance of the models was evaluated. This study's hold-out approach was chosen to compare the performances of the supervised models. In the hold-out approach, the data is divided into two parts, one of which is used to train the model and the other for validating and testing it. For the model performance evaluation, several measures can be used. This study calculated accuracy, error, and more comprehensive measures such as sensitivity, specificity, positive predictive value (PPV), and F-measure. These values should be close 1 For the C4.5 algorithm, additional measurements were also acquired like kappa statistics, mean absolute error (MAE), root mean absolute error (RMAE), relative absolute error (RAE), root relative squared error (RRSE).

Results

In this section, first, the results of the apriori algorithm are given based on all three-smoking behaviours. The rules were handled separately according to smoking behaviour. In this study, the threshold value was determined as 0.02. The top rules for each smoking behaviour are shown in Table 2, and their comments are presented just below.

Table 2: Main Association Rules for Smoking Behaviour

Smoking Behaviour	lhs	rhs	support	confidence	coverage	lift	count
Smokers	Gender=1, MaritalStatus=2, NewBMI=2, Alcohol=1, NewWorkingStatus2=1	NewSmoking=1	0.02001873	0.6909091	0.02897448	2.256450	342
Non-Smokers	Gender=2, DailyActivity=1, Alcohol=2, NewAge=1, NewWorkingStatus=4	NewSmoking=2	0.02417467	0.9582367	0.02522828	1.768638	413
Quit Smokers	Gender=1, MaritalStatus=2, NewAge=6,	NewSmoking=3	0.02089675	0.5368421	0.03892531	3.531540	357

For the smokers, the individuals/participants who are male, married, have a body mass index in the normal range, alcohol consumer, employer, self-employed or employees have been determined as smokers with a probability of 69%. Participants who provide this association are 0.02% of all participants. This rule has a lift value of 2.25, which is highly reliable.

Non-smokers, women in the 15-24 age range, who mostly sit or stand during the day, are non-alcohol users and do not smoke, with a probability of 95%. The incidence of this association among all participants is 0.02%. For quit smokers, it has been observed that married men in the 65-74 age range have quit smoking with a probability of 53%. Participants who provide this association together constitute 0.02% of all participants.

After all association rules were analyzed, it was shown that smokers tend to have characteristics like alcohol consumption, being male and married, mostly ages between 35-44 come to the forefront, whereas non-smokers tend to have characteristics like being woman, single, non-alcohol consumer, good general health, and ages between 15–24. For quit-smokers, married, elderly (65-74 years old), and retired men with low-daily activity are placed in this category.

For the visualization part of the association rules, the plot () function was used to illustrate the found association rules. The scatter diagram is the default approach used by the ArulesViz package's plot() function. The scatter diagram plots support and confidence intervals on the axes. As a third criterion, lift values can be used to depict the points by colours. The ruler showing the colour scale is located to the right of the graph. The graphic displaying the association rules for smokers is shown in Figure 1. The rules are represented by each point on the graph. The dark dots indicate the strength of the lift measure. Support values are between 0.05 and 0.15, while confidence values are between 0.2 and 0.9.

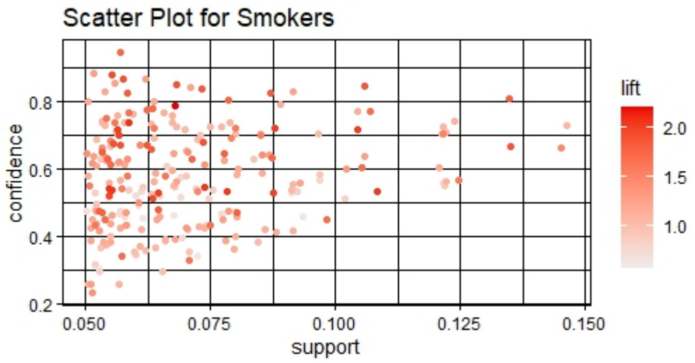

Figure 1: Scatter Plot For Smokers

Regarding the results of supervised machine learning methods, the data set was divided into 80% training, 20% test and 70% training, 30% test, according to the hold-out method. The analyses were repeated in both ways, and better performance results were obtained with the 80/20 split. Smoking behaviour was handled by two groups, smokers and non-smokers (quit smokers were included in non-smokers) to make better predictions. The findings are given in Table 3.

Table 3: Results of Model Performance Evaluation

Hold Out: 80/20		
	C4.5	**RF**
Accuracy	0.730	0.909
Error	0.269	0.090
Sensitivity	0.408	0.782
Specificity	0.873	0.965
PPV	0.587	0.908
F-Measure	0.481	0.840

In this section, at first, the top six model performance evaluation metrics obtained from the dataset are given respectively. When the metrics are examined, it has been seen that the RF model performs better than C4.5.

In addition to the results in Table 3, based on the C4.5 algorithm, 78.87% of the instances are correctly classified, whereas 21.13% are incorrectly classified. The mean absolute error of the model is 0.3138, the root mean squared error rate is 0.3961, the relative absolute error is 73.84%, and the root relative squared error is 85.93 %. Kappa statistics of the model are calculated as 0.4576.

Discussion and conclusion

Diseases caused by smoking are one of the most important issues that threaten human health around the world. Despite all the precautions taken to quit smoking by the various institutions and governments, thousands of people die every year due to smoking. Therefore, understanding the factors affecting this behaviour has gained great importance. On the other hand, over the past ten years, machine learning methods have been used to obtain some inferences from health-related data sets.

This study presents results from different machine learning methods to determine factors affecting the smoking behaviour of the respondents aged 15 and over based on the Turkey Health Interview Survey. First, the aim of implementing the association rule was to address and analyze the attitudes of individuals about smoking behaviour from different aspects. Therefore, the analysis includes socio-economic and health-related questions to detect associated smoking-related factors. According to the association rules, being male, being an alcohol consumer, being married, having a job and being in the normal range concerning BMI are the discriminative attributes for smokers. In contrast, respondents who are female in the 15-24 age range, non-alcohol users and continue their education are classified as non-smokers. The role of education, especially for females in this regard, is undeniable. This finding can be interpreted as education raising awareness in people about smoking. The age range 15-24 represents young people in the high school and university age group. In the quit-smokers group, respondents are male in the 65-74 age range and married. It shows that female individuals prioritize their health more by avoiding alcohol and smoking compared to males. Similarly, it can be said that retired individuals tend to abstain from smoking due to the health challenges they may frequently encounter in their retirement years. These findings of this study are consistent with the findings of Durmuşoğlu and Kocabey (2021), Dumortier,Beckjord,Shiffman,Sejdić, (2016), and Singh and Katyan (2019). This study highlighted that gender and alcohol use are outstanding factors in determining smoker individuals.

In the classification part of the analysis, C4.5 and RF algorithms were used to predict the smoking use of the people. In addition, the study compares the performances of the C4.5 and RF algorithms in predicting the smoking use of participants. To make better predictions, the dependent variable smoking behaviour was handled by two groups: smokers and non-smokers (quit smokers were included in non-smokers). Depending on this, better results were obtained from the RF algorithm than the C4.5 algorithm. Also, the RF model could predict an individual's smoking behaviour with an accuracy higher than 90%. Despite the high accuracy value obtained with the RF algorithm, this value can be improved using different algorithms. This result supports the findings obtained from previous studies conducted by Singh et al. (2022)

On the other hand, there have been some limitations concerning attributes used in the analyses. Some important questions could not be included, such as parents' smoking status and the resident type of the respondents in the analyses. It is thought that the research will make an important contribution to the literature as one of the examples showing how machine learning algorithms can be applied to the health-related data set obtained in Turkey.

Peer-review:

Externally peer-reviewed

Conflict of interests:

The authors have no conflict of interest to declare.

Grant Support:

The authors declared that this study has received no financial support.

Acknowledgement:

The authors thank the Turkish Statistical Institute for providing the Turkey Health Interview Survey 2019.

Author Contributions:

Idea/Concept/Design: **N**.T Data Collection and/or Processing: **N.T**. Analysis and/or Interpretation: **G.K.A**. Literature Review: **A.B.Ö.S**, Writing the Article: **N.T.**, **G.K.A.**, **A.B.Ö.S** Critical Review: **N.T**, Approval: **N.T.**, **G.K.A.**, **A.B.Ö.S**

References

Abo-Tabik, M., Benn, Y., & Costen, N. (2021). Are Machine Learning Methods the Future for Smoking Cessation Apps? Sensors, 21(13), 4254.

Abo-Tabik, M., Costen, N., Darby, J., & Benn, Y. (2019, August). Decision Tree Model of Smoking Behaviour. In 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (pp. 1746-1753). IEEE.

Coughlin, L. N., Tegge, A. N., Sheffer, C. E., & Bickel, W. K. (2020). A machine-learning approach to predicting smoking cessation treatment outcomes. Nicotine and Tobacco Research, 22(3), 415-422.

Ding, X., Yang, Y., Stein, E. A., & Ross, T. J. (2017). Combining multiple resting-state fMRI features during classification: optimized frameworks and their application to nicotine addiction. Frontiers in human neuroscience, 11, 362.

Dumortier, A., Beckjord, E., Shiffman, S., & Sejdić, E. (2016). Classifying smoking urges via machine learning. Computer methods and programs in biomedicine, 137, 203-213.

Durmuşoğlu, Z. & Kocabey Çiftçi, P. (2021). Socio-demographic determinants of smoking: A data mining analysis of the Global Adult Tobacco Surveys. Turkish Journal of Public Health, 19 (3), 251-262. DOI: 10.20518/tjph.884692

Evenhuis, A., Occhipinti, S., Jones, L., & Wishart, D. (2023). Factors associated with cessation of smoking in health professionals: a scoping review. Global Health Action, 16(1). https://doi.org/10.1080/16549716.2023.2216068

Genuer, R & Poggi, M. (2020) Random Forest with R. *Use R!* Springer.

Goodchild, M., Nargis, N., & d'Espaignet, E. T. (2018). Global economic cost of smoking-attributable diseases. Tobacco control, 27(1), 58-64.

Han, J. ve Kamber, M. (2006), *Data mining: concepts and techniques* (the Morgan Kaufmann Series in data management systems), 2nd Edition., Morgan Kaufmann Publishers, ISBN: 978-1-55860-901-3.

Issabakhsh M, Sánchez-Romero LM, Le TTT, Liber AC, Tan J. (2023) Machine learning application for predicting smoking cessation among US adults: An analysis of waves 1-3 of the PATH study. PLOS ONE 18(6): e0286883. https://doi.org/10.1371/journal.pone.0286883

Jitenkumar Singh, K., Jiran Meitei, A., Alee, N. T., Kriina, M., & Haobijam, N. S. (2022). Machine learning algorithms for predicting smokeless tobacco status among women in Northeastern States, India. International Journal of System Assurance Engineering and Management, 13(5), 2629-2639.

Koslovsky, M. D., Swartz, M. D., Chan, W., Leon-Novelo, L., Wilkinson, A. V., Kendzor, D. E., & Businelle, M. S. (2018). Bayesian variable selection for multistate Markov models with interval-censored data in an ecological momentary assessment study of smoking cessation. Biometrics, 74(2), 636-644.

Maginnity, J. D. (2020). Comparing the Uses and Classification Accuracy of Logistic and RF Models on an Adolescent Tobacco Use Dataset (Doctoral dissertation, The Ohio State University).

Mak, K. K., Lee, K., & Park, C. (2019). Applications of machine learning in addiction studies: A systematic review. Psychiatry research, 275, 53-60.

McCormick PJ, Elhadad N, Stetson PD. (2008) Use of semantic features to classify patient smoking status. AMIA Annu Symp Proc.; 450-454.

Nollen NL, Ahluwalia JS, Lei Y, Yu Q, Scheuermann TS, Mayo MS. (2016) 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.;18(4):386-394

RStudio, 2023, Home - RStudio, http://www.rstudio.com/, [Accessed: May 2023].

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22.

Singh J., Meitei K., J., A., Alee, N. T., Kriina, M., & Haobijam, N. S. (2022). Machine learning algorithms for predicting smokeless tobacco status among women in Northeastern States, India. International Journal of System Assurance Engineering and Management, 13(5), 2629-2639.

Singh, A., & Katyan, H. (2019). Classification of nicotine-dependent users in India: a decision-tree approach. Journal of Public Health, 27, 453-459.

Thakur, S. S., Poddar, P., & Roy, R. B. (2022). Real-time prediction of smoking activity using machine learning based multi-class classification model. Multimedia Tools and Applications, 81(10), 14529-14551.

TurkStat (2019), Turkey Health Interview Survey 2019

United Nations (2015). Transforming our World: The 2030 Agenda for Sustainable Development. https://sustainabledevelopment.un.org/post2015/transformingourworld/publication

WHO (2021). WHO report on the global tobacco epidemic 2021: addressing new and emerging products. https://www.who.int/publications/i/item/9789240032095.

WHO (2022). Noncommunicable diseases. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

WHO (2023) "Obesity and overweight", https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [Accessed: January 2023]

WHO. (2019). WHO global report on trends in prevalence of tobacco use 2000-2025. https://www.who.int/publications/i/item/who-globalreport-on-trends-in-prevalence-of-tobacco-use-2000-2025-third-edition

Zhang, Y., Liu, J., Zhang, Z., & Huang, J. (2019). Prediction of daily smoking behaviour based on decision tree machine learning algorithm. In 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC) (pp. 330-333).