A Machine Learning Approach for Predicting Nicotine Dependence

Mohammad Kharabsheh[1]
Computer Information System, The Hashemite University
The Hashemite University, Zarqa, Jordan

Omar Meqdadi[2]
Software Engineering, Jordan University of Science and Technology Jordan University of Science and Technology
Irbid, Jordan

Mohammad Alabed[3]
Bio-Medical Engineering, The Hashemite University
The Hashemite University, Zarqa, Jordan

Sreenivas Veeranki[4]
Medical Science, University of Texas Medical Branch
University of Texas Medical Branch, USA

Ahmad Abbadi[5]
MD, Tobacco Control Intern,
World Health Organization
Amman, Jordan

Sukaina Alzyoud[6]
Community & Mental Health Nursing,
The Hashemite University, Zarqa, Jordan

***Abstract*—An examination of the ability of machine learning methodologies in classifying women Waterpipe (WP) smoker's level of nicotine dependence is proposed in this work. In this study, we developed a classifier that predicts the level of nicotine dependence for WP tobacco female smokers using a set of novel features relevant to smokers including age, residency, and educational level. The evaluation results show that our approach achieves a recall of 82% when applied on a dataset of female WP smokers in Jordan.**

***Keywords*—Machine learning; nicotine dependency; Women; Waterpipe; classification**

I. Introduction

Bioinformatics is offered as a multidisciplinary field that helps researchers improving methods and software tools in order to understand biological data for several purposes under consideration by human beings. Bioinformatics is based on the employment of biology, computer science, mathematics and statistics to examine and interpret biological data. That is, bioinformatics is considered as a term for the body of biological researches that use computer programming and techniques as a part of their methodology. Additionally, the reference to some analysis "pipelines" is commonly used in the field of genomics.

Tobacco smoking is one of the most problematic public health problems that requests research, policy, and program initiatives [1]. Tobacco smoking behavior is related to several dimensions such as personal, social, community factors. Thus, knowing these factors is useful in developing a classification model for that behavior. For instance, solutions for waterpipe smoking cessation could be enriched by developing of models that classify smoker behaviors and thus may generate new hypotheses for clinical research. To do this, an accurate selection of variables that are strongly related to the behavior of smokers is required.

Unfortunately, to the best of our knowledge, there is not much support for automated generation of nicotine dependency levels from clinical datasets using machine learning classifiers. This study aims at determining whether machine learning approaches can assist in classifying waterpipe smoking behaviors of women. In this paper, we apply machine learning techniques to a corpus of women tobacco smoking questionnaires, which were previously developed by the authors of this work in [2], to discover and detect *Nicotine Dependence level* of Jordanian women. One of the long-term goals of the authors is to provide knowledge-rich environment that improves the quality of clinical decisions.

In particular, we formulate our study in the form of two research questions:

RQ1: Can we accurately predict if nicotine dependence level using classification factors?

Our obtained results show that we can build highly accurate prediction models for detecting women waterpipe smokers' level of nicotine dependence. For instance, we developed a machine learning classifier that achieves a recall of 81% recall and a precision of 75%.

RQ2: Which factors are the most important as predictors of nicotine dependence level?

We constructed a decision tree classifier and performed a Top Node analysis to identify the most informative factor for predicting nicotine dependence level. Nicotine dependence is a state of dependence upon nicotine [3].

The remainder of this paper is organized as follows. Section II reviews related. Section III discusses the methodology we follow in this study. Section IV presents the obtained results of our study. Section V discusses our obtained results. Section VI introduces the main threats to validity of

this work followed by Section VII with the conclusion of the paper and some plans for future research.

II. Related Works

The capability of using databases in order to extract useful information for quality health care is a vital for the success of healthcare institutions [4]. Historically, there are a several studies of investigations that necessity of using learning classifiers with Healthcare application and health informatics classifications.

Shouman et al. [5] performed nearest neighbor approach on benchmark data set to explore the performance of such approach in the diagnosis of heart diseases. Their approach achieves an accuracy of 97.4%. Brown et al. [6] apply SVMs on gene expressions in order to classify genes based on functionality. The obtained results show that SVMs are well performed with the problem of microarray gene classification. Nahar et al. [7] examined the effectiveness of classifier with predictive apriority for classifying heart disease in men and women.

The effectiveness of decision tree, neural network and naïve Bayes network in predicting heart attack were explored in [8]. The obtained results claim that the Naïve Bayes fared outperformed decision Trees as it could identify all the significant medical predictors. The study in [9] proposed and evaluates the effectiveness of a learning classifier on Pima Indian diabetes dataset. The results claimed that the machine learning is effective to detect diabetes disease diagnosis. The performance of neural network, Fuzzy logic and decision tree in diagnosing diabetes were examined in [10].

Kampourakia et al. [11] have introduced a web-based application that is based on SVMs to makes automatic diagnoses about health problems. Razzaghi et al. [12] have proposed multilevel SVM-based algorithms. They evaluate the proposed approach on public benchmark datasets with imbalanced classes and missing values in health applications. Their results show that multilevel SVM-based method produces accurate and robust classification performance. Yu et al. [13] present SVM approach to classify persons with and without common diseases. The approach shows effectiveness in detecting persons with diabetes and pre-diabetes in a sample of the U.S. population. The study in [14] proposed SVM approach that was trained using several terminological features to assign protein function and then choose passages based on the assignments.

To the best of our knowledge, this is the first work in the area of investigating the role of machine learning in detecting the nicotine dependence level of smoking women.

III. Proposed Methodology

In this section, we describe the design of our study. Initially, we introduce the dataset that is used in our study, and next we list the factors that are considered in our classification approach. Finally, we provide the prediction models and the performance metrics that is used in the evaluation of proposed models.

A. Studied Dataset

Our study was conducted among a sample of Jordanian women. A total of 108 women participated in the study with an age range of 18 to 56 years (mean = 26, SD± 9). Almost all participants 94.5% reside in an urban setting. More than two third 69.7% had a university degree. Thirty eight percent of study participants were students.

To produce concrete results, we collected our dataset over three points of time (two weeks before, two weeks into-, and two weeks after Ramadan. Objective measures were collected over three times before-, once during-, and after Ramadan. The study was conducted in the gynecology- obstetric clinics of two hospitals (one governmental and one private) in Amman city - Jordan. On average 35 patients are seen on daily basis at the clinics. All the gynecology- obstetric clinics of both hospitals were included to recruit none-pregnant study participants. In addition, all antenatal clinics affiliated with Hiba hospital were visited to recruit pregnant women. Inclusion criteria was women who are 18 years or older; able to read and write Arabic; absence of serious illness or being identified as high-risk patient.

The Women Tobacco Smoking Questionnaire (WTSQ): developed by the Principle Investigator, which is designed as a single measure to assess pattern of tobacco smoking among women. The questionnaire consist of four sections: (1) Demographics which includes age, educational level, marital status, etc., (2) tobacco smoking status asking about history of smoking habits, waterpipe smoking habits, (3) depression symptoms scale [15], this scale includes 6 items that assess the presence of depression, (4) Waterpipe Nicotine Dependence Scale [11], this scale measures level of nicotine dependence among waterpipe smokers, (5) waterpipe smoking during Ramadan, this part ask participants about their waterpipe smoking during Ramadan. Response options in the questionnaire vary based on the construct and items measuring that construct. They range from Likert-type responses; yes/no responses, fill in the blank, to a multiple-choice question.

B. Classification Factors

To classify and predict women waterpipe smokers' level of nicotine dependence, we considered 19 factors as shown in Table I. We use these factors since they perform well in traditional tobacco prediction research and represent standard factors for the desire for women to smoke waterpipe [2, 16]. Another rational is that these factors also cover experiencing cravings that leads them to smoke waterpipe [17, 20]. It was demonstrated that young initiation age of smoking linked to being a regular smoker at a later age [17, 18, 19]. Moreover, it was shown that number of tobaccos smoked significantly related to level of nicotine dependence [17, 21, 25, 29, 30].

C. Creating the Corpus

The primary step involved in performing our classification purpose is creating the corpus that represents the input of machine leaning classifiers. For this work, the corpus includes the extracted values relevant to every classification factor for each instance of our studied dataset. These values are extracted from the women's responses of WTSQ.

TABLE I. SUMMARY OF CLASSIFICATION FACTORS

Classification Factor	Definition
Age	Participants age should be 18 years or older
Residence	place of residence
Education-Level	The highest level of education of the participants.
Work-Status	Type of work
Current-Tobacco	Number of tobaccos of any type smoked in the last 30 days even if one puff
Past-Tobacco	Have you smoked tobacco in the past
Tobacco-Age	How old were you when you first started smoking tobacco
Tobacco-Type	Type of tobacco products currently smoke
Number-of-Cigarette	Number of cigarettes smoked
Current-Waterpipe	Number of waterpipe of any type smoked in the last 30 days even if one puff
Waterpipe-Week	Number of waterpipe of any type smoked in the last seven days even if one puff
Waterpipe-Inhale	Do you inhale the smoke when you smoke waterpipe
Waterpipe-Stop	How many times could you stop waterpipe for more than 7 days
Waterpipe-Month	Same as current tobacco smoke
Waterpipe-Alone	Do you smoke waterpipe alone
Waterpipe-Need	Have you ever felt that you actually need to smoke waterpipe
Waterpipe-Income	The percentage of income that you regularly spend for waterpipe smoking
Waterpipe-Days-Without	Number of days you could spend avoiding waterpipe
Waterpipe-Cigarette-Instead	When I feel the need to smoke waterpipe and it is not available, I smoke a cigarette instead

Next, we label each instance of the corpus with the associated nicotine dependency level by finding the summation of all participant answers and divided to three groups (A-High Score, B, C) with equal participants in each group. Table II summarizes the corpus information.

D. Prediction Models and Evaluation Metrics

There are numerous machine learning techniques such as Support Vector Machines (SVM) that can help the achievement of our classification goals. In this study, we chose to use the below classification approaches, which have been used with relative success in prior classification work [5, 22, 23, 27] with different domains and problems.

TABLE II. SUMMARY OF CORPUS INFORMATION

Total number of instances	Number of Level A	Number of Level B	number of Level C
108	36	36	36

- Support Vector Machine Learner (SMVL): is an approach that increases the dimensionality of data until the data points are differentiable in some dimension.
- Bayesian Learner (Naïve Bayes): is a Bayesian learner, which is like the techniques that are used in classifying email spam.
- K-Star: is a nearest neighbor algorithm that utilizes a distance metric such as the Mahalanobis distance.
- IBk: is a single-nearest-neighbor algorithm, which classifies data entities via using the closest associated vectors in the training set through distance metrics.

Several good quality implementations of SVM are available. We use the WEKA toolkit implementation [21, 31] to build our model.

With supervised classifiers, the dataset is divided into two sets: a training set and a test set. The training set is used to train the classifier, while the accuracy of the model is measured using the test set. In our study, the decision of which subset is used as a training set or a test set is controlled by 10-fold cross-validation [23] technique, which was widely used.

We used four performance metrics to evaluate the efficacy of each classifier: Precision that represents the percentage of retrieved instances that are relevant (P = True Positives / (True Positives + False Positives)). Recall represents the percentage of relevant instances that are retrieved (R= True Positives / (True Positives + False Negatives)). F-Measure is metric that is calculated by a combination of precision and recall ((2 * R * P) / (R + P)), and thus its value is between 0 and 1. ROC represents the area under the Receiver Operating Characteristic (ROC) curve, which is based on the plotting of true positives versus false positives.

IV. STUDY RESULTS

We now present the details behind the results and the outcomes of our study that were obtained by answering our research questions posed before.

A. RQ1: Can we Accurately Predict Nicotine Dependence Level using Classification Factors?

To answer this question, we want to build prediction models to help classifying women waterpipe smokers' level of nicotine dependence, and we want to know if we can accurately predict these dependncies using the factors that we examined early.

As we mentioned earlier, we used several SVM approaches to build our prediction models. Also, we used the 10-fold cross validation approach to divide our inputted dataset into training and test sets. The effectiveness of our models is evaluated using the recall, precision, ROC, and F-measure metrics. Now let us look at our proposed classifiers that were trained using a combination of all factors that are given in Table I. The performance results of our classifiers are shown in Table III.

It is observed that our obtained results show the prediction improvement when comparing our developed classifier with the baseline model in terms of all evaluation measures. For instance, when comparing our SMO classifier to the baseline

model, the improvement ratio is 0.82 in terms of recall and 0.43 in terms of precision. *That is, we can build highly accurate prediction models for detecting women waterpipe smokers' level of nicotine dependence*. Thus, our results demonstrate that several factors of a person such as age, level of education, and the number of cigarettes impact the nicotine dependence level of Jordanian women.

Our second observation is that SMO and Naïve Bayes offer better classification accuracy than the rest of the machine learning classifiers. For instance, Naïve Bayes computes a probability for each class based on the probability distribution in the training dataset. Therefore, with each training example, the prior and the probability can be updated dynamically to achieve flexibility and robustness to classification errors. On other hand, the SMO learner achieves better accuracy because of increasing the dimensionality of data until the data points are differentiable in some dimension. Additionally, the space usage needed for SMO is linear in the size of training set; therefore it allows SMO to handle very large training sets with higher accuracy.

B. *RQ2: Which Factors are the Most Important as Predictors of Nicotine Dependence Level?*

Here, we try to evaluate the performance of different factor group combinations for performing our classification. To do this, we combined related factors into four groups, as follows:

- Group1. Age, Residence, Level_Education, Work Status, Waterpipe-Income
- Group2. Tobacco Past, Tobacco Age, Tobacco-Type, Number-of-Cigarette
- Group3. Current-Waterpipe, Waterpipe-Week, Waterpipe-Month, Waterpipe-Need, Waterpipe-Cigarette-Instead
- Group4. Waterpipe-Inhale, Waterpipe-Stop, Waterpipe-Alone, Waterpipe-Cigarette-Instead

To answer RQ2, a classification model is trained using factors from each factor group and then its precision and recall are measured. We developed these classification models using the SMO approach since, as we discussed early, it has outperformed other classification approaches in term of recall and precision.

Group4 produced poor results, see Table IV. One reason could be contributed to the fact that waterpipe mostly smoked in gatherings and not alone. Another interpretation could be that women did not think to stop waterpipe smoking since previous studies [24] showed that they do not perceive it as harmful to health. Group1 produces the best results. One explanation might be as indicated in previous findings that nicotine dependence increases with age. Additionally, living in urban and sub-urban settings could facilitate smokers' access to places that serve waterpipe or sell waterpipe tobacco. Working and having personal income could enable individuals to be economically independent to spend money on waterpipe smoking.

In an attempt to get a zoomed-in picture, we also evaluate the effectiveness of each factor independently as a predictor of nicotine dependence level. Instead of measuring the performance of each factor in predicting nicotine dependence level, we chose to use a decision tree to rebuild a classifier that is trained using all classification factor given in Table I.

The essential algorithm that builds the decision tree is the C4.5 algorithm [26]. C4.5 follows the greedy divide and conquer approach using the training data, where it begins with an empty tree, and then it adds decision nodes (leaf) at each level. Moreover, the information obtained using a specific factor/attribute is calculated, and then the attribute with the highest information gain is chosen. Additional analysis is performed to determine the threshold (e.g., cut-off) value at which to split the attribute. This process is recursively repeated at each level until the number of records in the leaf reaches the specified threshold.

With decision trees, we could perform the Top Node analysis [28] to order factors based on their effectiveness in predicting nicotine dependence level. The Top Node approach examines the structure of a decision tree [18], and counts the appearance of each factor at each level of the tree. Then, the importance rank of each factor is determined by the combination of the tree level in which the factor appears and the occurrence count of the factor. That is, the root node of the decision tree represents the most important factor and so the factors become less important as we move down the tree. The performance of our decision tree classifier that was trained using the combination of all factors and was build using the C4.5 algorithm is given in Table V. As we could observe, SMO and Naive Bayes have produced better results than decision tree.

On the other hand, the results of the Top Node analysis are shown in Table VI. The table shows the top factors that appear in the first three levels (e.g., levels 0, 1, and 2) of the created tree along with number of occurrences of each top factor. *For our dataset,* the age factor is the most influential than other considered factors. This finding could be contributed to the assumption that women have more ability to smoke water pipe more freely with age. Interestingly previous studies demonstrated that nicotine dependence level increase with age [15]. Moreover, we would assume that with age it becomes more difficult for women to decrease or quit smoking.

TABLE III. CLASSIFICATION RESULTS OF NICOTINE DEPENDENCE DETECTION USING MACHINE LEARNING ALGORITHMS

Learner	Recall	Precision	F-Measure	ROC
SMO	0.82	0.43	0.56	0.90
Naïve Bayes	0.79	0.42	0.55	0.91
K-Star	0.47	0.21	0.29	0.69
IBk	0.64	0.29	0.41	0.74

TABLE IV. CLASSIFICATION RESULTS OF NICOTINE DEPENDENCE DETECTION USING COMBINATION OF FACTORS

Group	Recall	Precision	F-Measure	ROC
Group1	0.75	0.38	0.51	0.84
Group2	0.63	0.28	0.39	0.74
Group3	0.45	0.16	0.24	0.65
Group4	0.42	0.15	0.22	0.61

TABLE V. CLASSIFICATION RESULTS OF NICOTINE DEPENDENCE DETECTION USING C4.5 ALGORITHM

Learner	Recall	Precision	F-Measure	ROC
C4.5	0.69	0.33	0.45	0.76

TABLE VI. RESULTS OF TOP NODE ANALYSIS FOR THE DECISION TREE ALGORITHM

Level	Frequency	Attribute
0	7	age
	3	Work-Status
1	9	Tobacco Past
	6	Level_ Education
2	12	Tobacco Current
	9	Tobacco-Type
	2	Residence

V. DISCUSSION ON FINDINGS

In this study, we have used the supervised classifiers to develop our approach. We showed the effectiveness of our classification approach in predicting women Waterpipe smoker's level of nicotine dependence. Our results provide the performance of the studied factors and attributes. The results suggest that our approach would help researchers in the planning for health management of female smokers.

We could conclude that the developed model outperforms a random guessing approach that would result in an overall misclassification rate. That is, comparing with random guessing would verify the strength of our model. We correctly achieved a recall of 82% and a precision of 43%. However, the produced model is based on a dataset that is extracted from answers of female smokers, and thus it is possible that includes false negative answers. Such sampling may be subjective and so could affect the classification performance. Therefore, we do not claim that our evaluation is without faults. Moreover, although our model achieves higher recall, the model dos not achieve high precision values. This could be a main weakness of our model since it represents a troubling finding. Other weaknesses are discussed as threat to validity in the next section. Possible future work could study how selection of study dataset of habits impact the ability of our approach, and study how to improve the precisions of our model by dealing with possible threats of validity of this study.

The research was undertaken using the supervised classifiers with specific classification approaches such as Bayesian Learner and decision trees. This research could be undertaken using other classification approaches or through the usage of unsupervised classifiers. However, we believe that unsupervised classifying model could be much difficult to understand and use in practice for health care planning where we are looking for simple and basic rules that practitioners could use. Also, it is shown in the literature that the used approaches, such as decision trees, outperforms other supervised classification approaches [28].

VI. THREATS TO VALIDITY

We now examine threats. We use datasets of 108 Jordanian women age range of 18 to 56 years, and thus it might not be representative of all women out there. There may be other factors that we did not consider in our work, such as family and friends waterpipe smoking, waterpipe smoking sessions, waterpipe smoking heads, and psychological status such as depressive modes. We plan to evaluate the effectiveness of other factors and dimensions in future.

In this work, we used several commonly used machine learning techniques such as support vector machine learner and decision trees. However, each of these techniques has its own limitations that could affect the validation of our obtained results. More research using other techniques might be part of our future work.

VII. CONCLUSION AND FUTURE WORK

The current study exploited the effectiveness of machine learning techniques in classifying and predicting nicotine dependence level of waterpipe smoking women. We have performed a study based on a set of factors obtained from a dataset of 108 women with an age range of 18 to 56 years.

This work presents machine learning classifiers based on support vector machine, Bayesian learner, nearest neighbor algorithm, and decision trees for predicting nicotine dependence level of Jordanian women. To build our models, we used a set of factors such as age, level of education, and working status.

Our results show that the presented prediction models have reasonable accuracy with 82% recall in the best case and 47% recall in the worst case. In addition, a precision of 43% is achieved in the best case and 21% in the worst case. Top Node analysis shows age is the most important factor in our classification.

We aim to explore more classification factors and study the effectiveness of other machine learning techniques in predicting nicotine dependencies in future studies in order to achieve better prediction performance. We plan to enrich our study by investigating more varies datasets from different countries and environments.

ACKNOWLEDGMENT

The current study was funded by the Deanship of Scientific Research at the Hashemite University. We would also like to thank Hiba Privet Hospital for approving the study and providing access to their out-patient's clinics.

REFERENCES

[1] WHO. (2018). WHO global report on trends in prevalence of tobacco smoking 2000–2025 (Second Edition ed.). Geneva: World Health Organization.

[2] Alzyoud, S., Veeranki, S.P., Kheirallah, K., Shotar, A.M., Pbert, L. (2016). Validation of the Waterpipe Tolerance Questionnaire among Jordanian School-Going Adolescent Waterpipe Users. Global Journal of Health Science 8(2)· 8(2):198-208. doi: 10.5539/gjhs.v8n2p198.

[3] D'souza, M. S., & Markou, A. (2011). Neuronal mechanisms underlying development of nicotine dependence: implications for novel smoking-cessation treatments. Addiction science & clinical practice, 6(1), 4.

[4] Eapen, A. G. (2004). Application of Data mining in Medical Applications. Ontario, Canada, 2004: University of Waterloo.

[5] Shouman, M., Turner, T., & Stocker, R., "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients," International Conference on Knowledge Discovery (ICKD 2012), 2012.

[6] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D., "Knowledge-based Analysis of Microarray Gene Expression Data using Support Vector Machines". Proceedings of the National Academy of Sciences, 2007.

[7] Nahar, J., Imama, T., Tickle, K., Chen, Y., "Association rule mining to detect factors which contribute to heart disease in males and females," Elsevier," 2013.

[8] Ms. Ishtake S.H ,Prof. Sanap S.A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research,2013.

[9] S. W. Purnami, A. Embong, J. M. Zain and S. P. Rahayu, "A New Smooth Support Vector Machine and Its Appli- cations in Diabetes Disease Diagnosis," Journal of Computer Science, Vol. 5, No. 12, pp. 1006-1011.

[10] Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013.

[11] A. Kampourakia, D. Vassisa, P. Belsisb, C. Skourlasa," e-Doctor: A Web Based Support Vector Machine for Automatic Medical Diagnosis," The 2nd International Conference on Integrated Information..

[12] Talayeh Razzaghi, Oleg Roderick, Ilya Safro, Nick Marko," Fast imbalanced classification of healthcare data with missing values," 18th International Conference on Information Fusion (Fusion), 2015.

[13] Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, Muin J Khoury, " Application of support vector machinen modeling for prediction of common diseases: the case of diabetes and pre-diabetes," BMC Medical Informatics and Decision Making2010.

[14] Rice SB, Nenadic G, Stapley BJ: Mining protein function from text using term-based support vector machines. BMC Bioinformatics. 2005

[15] Haddad, Linda G., Ali Shotar, Janet B. Younger, Sukaina Alzyoud, and Claudia M. Bouhaidar. "Screening for domestic violence in Jordan: validation of an Arabic version of a domestic violence against women questionnaire." International journal of women's health 3 (2011): 79.

[16] E. Aboaziza and T. Eissenberg, "Waterpipe tobacco smoking: what is the evidence that it supports nicotine/tobacco dependence?," *Tobacco Control*, vol. 24, no. Suppl 1, pp. i44-i53, 12/09,09/16/received,11/20/accepted 2015.

[17] R. Bahelah et al., "Waterpipe smoking patterns and symptoms of nicotine dependence: The Waterpipe Dependence in Lebanese Youth Study," Addictive Behaviors, vol. 74, pp. 127-133, 2017/11/01/ 2017.

[18] W. Maziak, K. D. Ward, and T. Eissenberg, "Factors related to frequency of narghile (waterpipe) use: the first insights on tobacco dependence in narghile users," Drug & Alcohol Dependence, vol. 76, no. 1, pp. 101-106, 2004.

[19] Alzyoud, S., Kheirallah, K. A., Weglicki, L. S., Ward, K. D., Al-Khawaldeh, A., & Shotar, A. (2014). Tobacco smoking status and perception of health among a sample of Jordanian students. International journal of environmental research and public health, 11(7), 7022-7035.

[20] R. A. Auf, G. N. Radwan, C. A. Loffredo, M. El Setouhy, E. Israel, and M. K. Mohamed, "Assessment of tobacco dependence in waterpipe smokers in Egypt," The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, vol. 16, no. 1, pp. 132-137, 2012.

[21] D. Mays, K. P. Tercyak, K. Rehberg, M.-K. Crane, and I. M. Lipkus, "Young adult waterpipe tobacco users' perceived addictiveness of waterpipe tobacco," Tobacco Prevention & Cessation, vol. 3, no. December, 2017 2017.

[22] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016 Oct 1.

[23] Efron, B., " Estimating the error rate of a prediction rule: improvement on cross-validation ", Technical Journal of the American Statistical Association, vol. 78, no. 382, pp. 316–331, 1983.

[24] Akl, E.A.; Jawad, M.; Lam, W.Y.; Co, C.N.; Obeid, R.; Irani J. Motives, beliefs and attitudes towards waterpipe tobacco smoking: a systematic review. Harm Reduct J. 2013, 10(1), 12, doi: 10.1186/1477-7517-10-12.

[25] Martinasek, M.P.; McDermott. R.J.; Martini, L. Waterpipe (hookah) tobacco smoking among youth. Curr Probl Pediatr Adolesc Health Care. 2011, 41, 34–57. doi:10.1016/j.cppeds.2010.10.001.

[26] Quinlan, J., " C4.5: programs for machine learning", Morgan Kaufmann Publishers Inc., 1993.

[27] Garcia, H., Shihab, E.," Characterizing and predicting blocking bugs in open source projects," In Proceedings of the 11th Working Conference on Mining Software Repositories (MSR'14), New York, NY, USA, pp. 72 - 81, 2014.

[28] Hassan, A., Zhang, K.," Using decision trees to predict the certification result of a build," In Proceedings of the 21st IEEE/ACM International Conference on Automated Software Engineering (ASE '06), pp. 189–198, 2006.

[29] Kheirallah, Khalid A., Sukaina Alzyoud, and Kenneth D. Ward. "Waterpipe use and cognitive susceptibility to cigarette smoking among never-cigarette smoking Jordanian youth: analysis of the 2009 Global Youth Tobacco Survey." Nicotine & Tobacco Research 17.3 (2014): 280-284.

[30] Valdivia-Garcia H, Shihab E, Nagappan M. "Characterizing and predicting blocking bugs in open source projects". Journal of Systems and Software. 2018 Sep 1;143:44-58.

[31] https://www.cs.waikato.ac.nz/ml/weka/.