Using machine learning algorithms to study the smoking behavior of Iraqi students

Raya N. Ismail

Department of Computer Sciences, College of Computer & math. Science, Tikrit University, Iraq
Raya_computer@tu.edu.iq

ABSTRACT

Many Machine learning studies analyzed smoking behavior through the relation of this complex behavior and smokers inherited genes other studies tried to predict the negative impact of smoking by monitoring body movements (like arm movement, breathing patterns). All these studies have limited ability to analyze the reasons led to this behavior specially for teenagers. In this study we present a methodology with five predictors to analyze the smoking behavior of school students in Iraq. The data obtained from National Youth Tobacco Survey Data set (NYTS) in 2019 which contain self-reported questions for 2560 individual arranged into 99 attributes. Naïve Bayes (NB), K*, PART, Logitboost, and REPTree, have been used to analyze the student smoking status from three vectors the overall smoking behavior, smoking cessation, and school environment impact.

The results showed good outcomes for the three vectors, K* model seems to be the best predictor for overall smoking behavior with 91,02 accuracy, Logitboost scored 84.6611 accuracy for smoking cessation, and 78.383 as best result in evaluating school influence. These results proof that machine learning models have promising ability to predict student smoking behavior.

Keywords: machine learning, smoking behavior, self-reported questions

1. Introduction

Smoking can classified as a main criteria that effect the humane body system with negative outcomes causing many harmful damages which is related to serious diseases like lung cancer specially for heavy smokers [1]. IMU sensor used to monitor and analyze smoking activity for workers in their working environment as daily activity like walking, sleeping and running to help healthcare professionals [2]. According to the reports of Health Care Organization smoking can cause death to 8.3 Million around the world by 2030 so it became necessary to give efforts in preventing tobacco smoking by analyzing the two main factors that controls this behavior, genetic factor and environmental factors [3]. Machine learning algorithms used to study the changes in human body caused by smoking, researchers find that addiction to tobacco changes the brain signals and the nicotine level in different cigarettes effects the human brain [4].

Sensors technology with wireless network played great role in monitoring and detecting harmful activities like smoking. Smart cellphones and smart watches used to capture the motion context to predict smoking events [5].

Since it is difficult to measure the levels of isoform expression and the changes in gene-level expressions caused by smoking cigarettes, however the use of isoform algorithm with RNA-seq showed that smoking is responsible of widespread isoform changes and usage of exon [6].

More than one third smokers tried to quit once yearly at least but 70% of these attempts fail. Many researches produced to understand why smokers can`t just stop smoking and what are the difficulties they are facing. Smoking status self-reporting method proved to be good way to estimate status of smokers and treatments [7].

In the field of wearable systems PACT2.0 life-activity monitoring devices designed to capture all life-activities related to smoking like the movement of hand, chest and lighter using Cigarette Tracer sensors these sensors build to detect hand to mouth movement, breathing patterns, and duration between lighting events. Laboratory tests applied to evaluate the accuracy of PACT2.0 data, then a computer analysis take place using 98% of the collected data, PACT2.0 proved to be good platform to study smoker behavior in their environment [8].

Machine learning Studies find a relationship between smoking activity and aging rates, female smokers may be twice older than nonsmoker chronological age, and male smokers predicted to be one and a half older [9].

Sensor signals from a wearable non-invasive system tested with supervised learning algorithm SVM as features to study smoke inhalations [10].

Beside death, diseases and changes in brain signals smoking can cause damages to pregnant women like congenital anomalies and miscarriage [11]. American studies stated that Tobacco smoking can be main factor to bladder cancer for both male and female smokers in the united states [12].

Risks related to smoking can be measured by the evaluation of HRV (hart rate variability) the experiments applied to 17 volunteers (smokers, non-smoker, ex-smoker) and results showed small changes in HRV for those how use nic-packs with amount of nicotine up to 6 mg [13].

Number of e-cigarette smokers in US youth increased despite the harmful impact on health. Factors like mental, social, and environmental determined by researchers to build machine learning models that can predict nicotine addiction of 6511 case[14].

The rest of the study arranged as follows: section two shows the enrolled data and the methods used to train and test data, section three presents the resulted performance of all used predictors, section four discusses these results and section five produce conclusions.

2. Methods and Data

2.1 Data Set

Data collected from Iraqi schools between 2010 and 2019 located in National Youth Tobacco Survey Data set (NYTS). Contained 2560 instances, 99 features. Each student in this Survey answered group of questions to analyze the impact of surrounding circumstances led to smoking in teenagers. We ignored the missing values and the non-smokers instances.

A numeric value represents the answer of a final weight calculated based on these answers. Ranges of data and characteristics showed in Table 1.

Table 1 Students characteristics

Group	Range	Highest percent
Age	(11-17) years	26.9% (14)
Sex	(male-female)	59.7% (male)
Grade	(1st middle-3rd middle)	40.4% (1st middle)
First cigarette age	(7-16) years	7.6% (12 or 13)
# cigarette per day	(1-20 or more) cigarette	3.9% (1 per day)
Smoking places	Places	9.1% (at home)
Seeing parent smoke	(every day-never)	25.1% (about every day)

2.2 Preprocessing steps

First we eliminate the missing values and enroll 2204 answers as showen in Fig.1.

The model applied with sets of features contains some students characteristics and a questionnaire like 1: number of cigarette in day? 2: age of first cigarette? 3: time between two cigarette? And other questions. To get best prediction results the features evaluated by

three types of models and the conjunction between the resulted sets used in the final evaluation step.

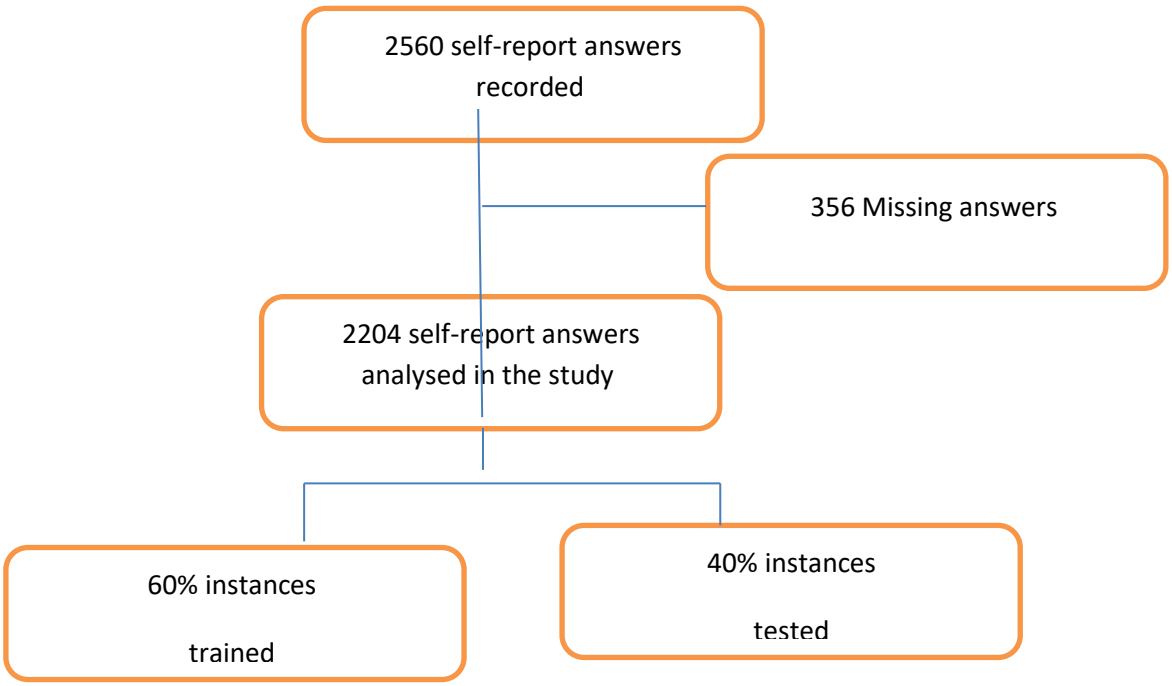

Fig1. enrolled data

To evaluate the overall student smoking behavior (smoker, ex-smoker, quitter) based on his answers so all attributes involved and apply feature selection ranker (Relife) (2)InfoGainRatio and (3)GainRatio to give the best result in predicting the relation between students answers and the percent of addiction in the Iraqi schools. Too many attributes may confuse the predictor and decrease the precision that is why we choose to rank and choose the best ranked attributes ,

In second experiment to predict if the student has attention to quit smoking or not the features related to smoking cessation involved.

Last experiment measures the school environment impact on students smoking potential so the every question joins between smoking and school is enrolled.

2.3 Machine learning Models

In this study five models used to analyze the student smoking behavior using self-report questions related to smoking circumstances. These models are: Naïve Bayes (NB), K*, PART(C4.5), LogitBoost, and RandomSubSpase (REPTree).

Naïve Bayes classifier first applies SupervisedDiscreetization to converts the

numeric features to nominal and uses the analyzed trained data to choose the values of precision estimator. We applied K* algorithm with the outcome of (CtsSubseEval+BestFirst) that generated a subset of the most relevant attributes then the model classifies the data as an instance-based method, which consists of labeling the test instances using similarity with the labels of training ones based on entropy-based equation. The PART model designs decision list for every iteration and a rule generated for a best leaf.

LogitBoost build with regression learner to classify binary classification problem, for the search process the class builds decision stump that connected with boosting algorithm. RepTree deigns a decision tree model to get best accuracy from the training data and improves complexity as it enhances the generalized accuracy. RepTree constructs more than one tree and selects pseud randomly feature subsets to build trees with random chosen sub spaces.

2.4 performance of evaluation models

the data divided to 6% training, 40% testing evaluate each model performance through statistical factors like (kappa statistic), number of correct labeled data (accuracy%), Mean absolute error, and the time taken to build the model with 10-folds cross-validation.

3. Results

Each predictor evaluated by three vectors accuracy, kappa statistic, and mean absolute error. Where the accuracy is one metric value to measure the classification models performance. Informally, accuracy is to estimate the degree our model correctness. Formally, accuracy has the following definition:

Accuracy= $\#\ of\ correct\ predictions / \#\ of\ total\ predictions$

[15]

However, Kappa Cohen's K-coefficient used to measure agreement degree of two variables by comparing the probability of agreement if the ratings are independent. Its ranges lies between -1 and 1 where the standard for acceptable kappa is arbitrary [16].

Mean absolute error is obtained by calculating the difference between the predicted values and the original values then calculate the average. To measure how far the predicted values from the correct output.

Yi^ is the predicted value if the i_{th} is a sample from n samples, and y_i is the correct value, so the mean absolute error (MAE) defined as the following:

MAE(y,y^)=$1_n \sum i = 0n - 1|yi - y^i|$

[16]

3.1 Smoking behavior predictions

To evaluate student overall smoking behavior each model trained then tested with ranked features, outcomes recorded to compare the values of Accuracy, Kappa Statistic, Mean Absolute error of the five models as shown in Table 2. Accuracy comparison illustrated in fig2.

Fig3. Shows the five models evaluation by kappa statistic, and fig4. Shows the MAE .

Table 2. Evaluation of machine learning models for overall behavior

Models	Accuracy	Kappa Statistic	Mean Absolute Error
NB	73.1641	0.7227	0.0159
K*	91.0200	0.9084	0.0725
PART	87.1875	0.8698	0.0023
LogitBoost	87.5000	0.8729	0.0022
REPTree	86.1719	0.8594	0.0067

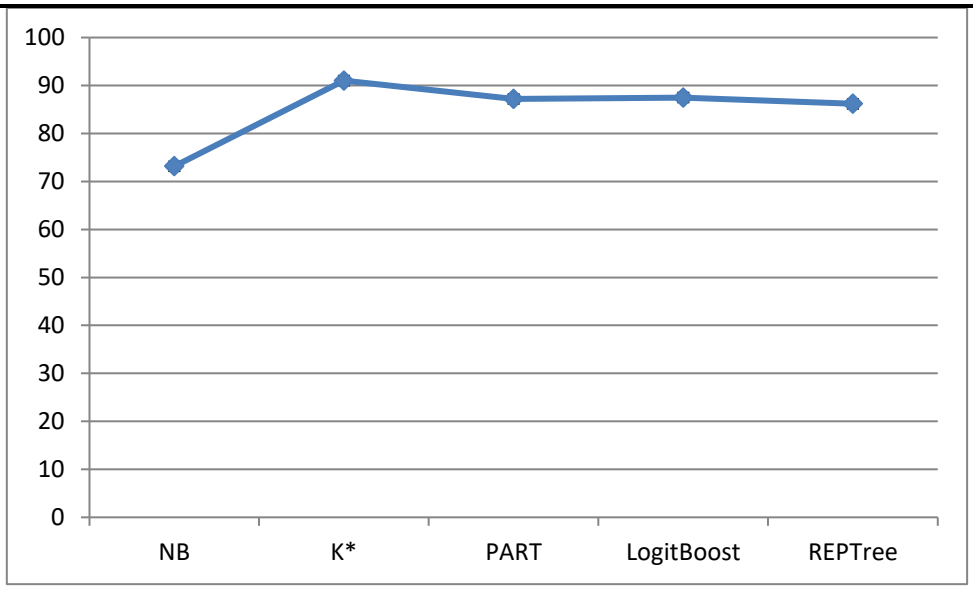

Figure 2. models accuracy for overall smoking behavior

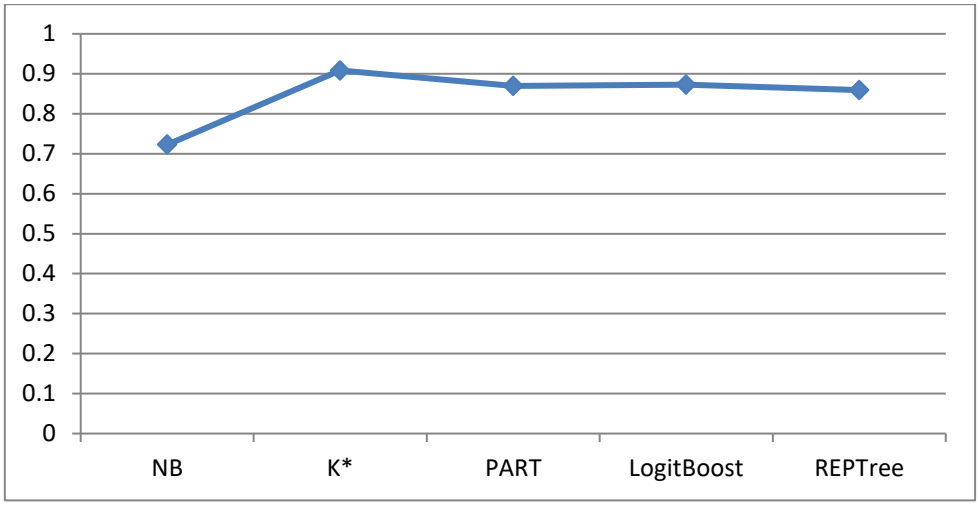

Figure 3. models kappa statistic for overall smoking behavior

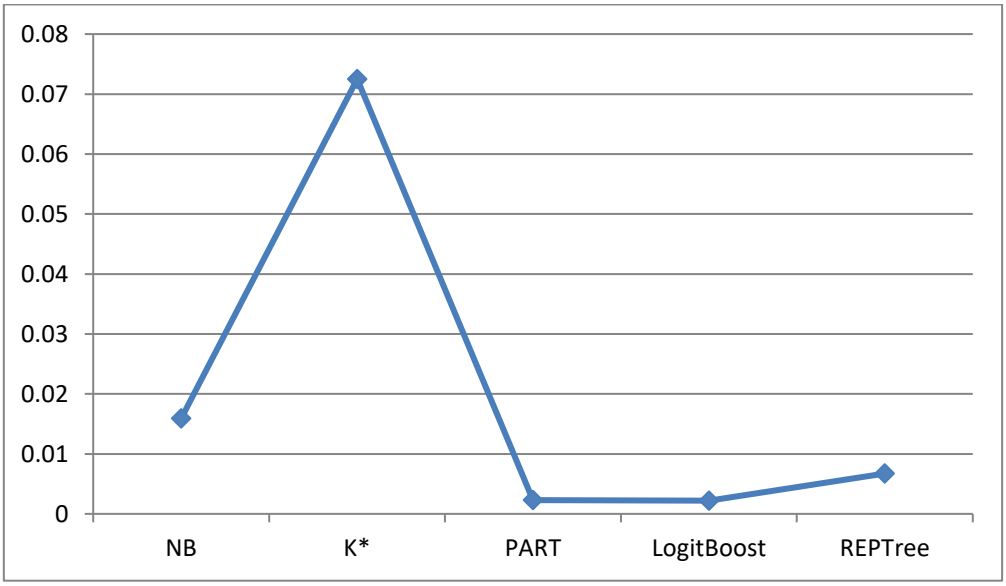

Figure 4. models MAE for overall smoking behavior

To evaluate the performance of the predictors we compared the time taken to build each one with 10 folds cross-validation. As shown in fig5.

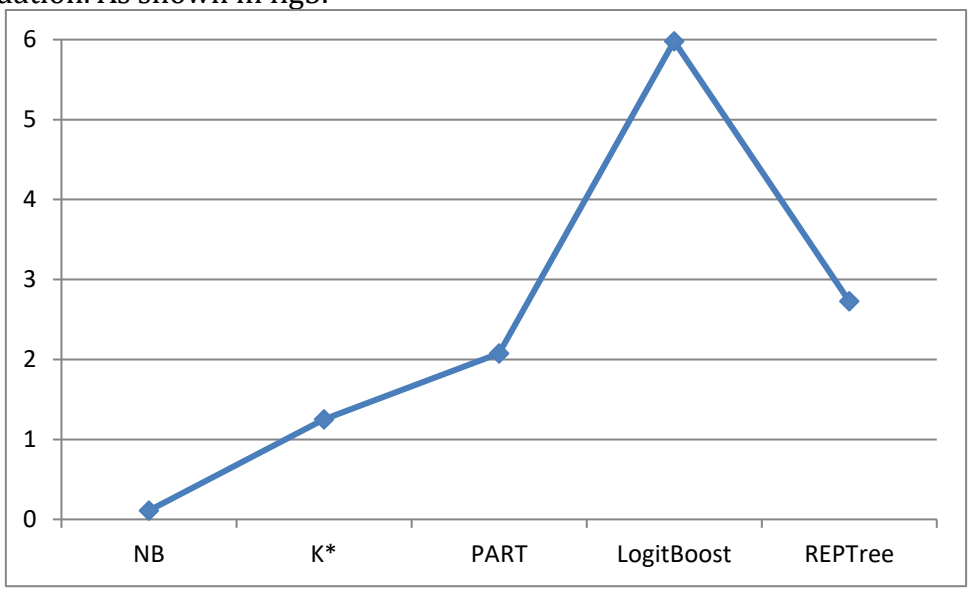

Figure 5. Time taken to build models for overall smoking behavior (in seconds)

3.2 Prediction of smoking cessation

In order to predict if the student intend to quit smoking based on his behavior that we can analyze from the self-report questions like: if he knows the consequences of smoking? if he tried to stop smoking now or during the past year?, if he received an advice or help from (friend, family, program, professional) to stop smoking?, if had educated in his classes about the danger of using tobacco products? etc.

The evaluation of the five models with smoke cessation self-reported questions showed in Table 3. The accuracy illustrated in fig 6.

A comparison of models kappa statistic presented in fig 7. And MAE in fig 8. For students smoking cessation.

Table 3. Evaluation of machine learning models for smoking cessation

	AC	KS	MAE
NB	82.9415	0.6439	0.0944
K*	80.0634	0.5627	0.1066
PART	82.0055	0.6330	0.1317
LogitBoost	84.6611	0.6819	0.1110
REPTree	82.2037	0.6330	0.1317

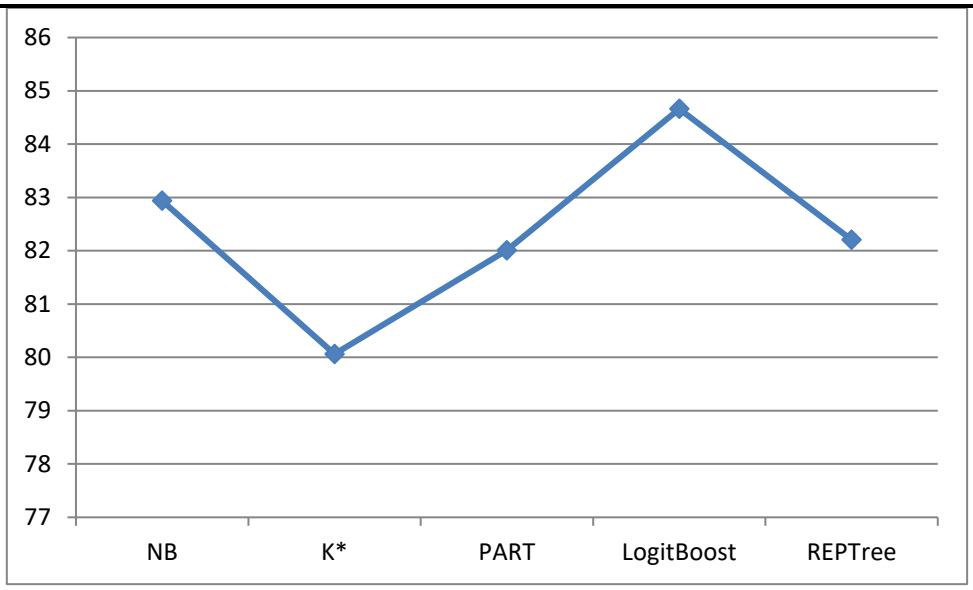

Fig 6. Models accuracy for smoking cessation

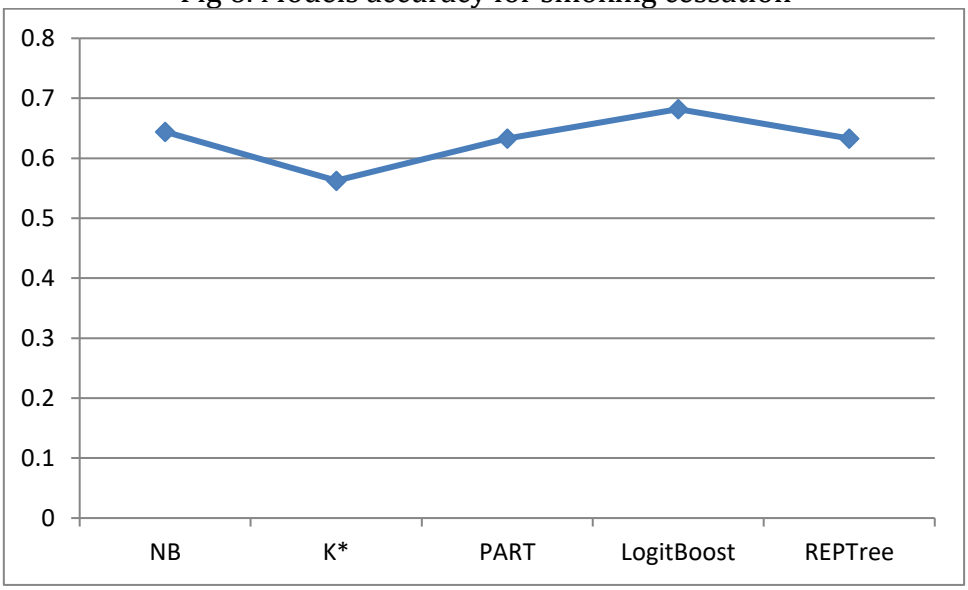

Fig 7. Models Kappa statistic for smoking cessation

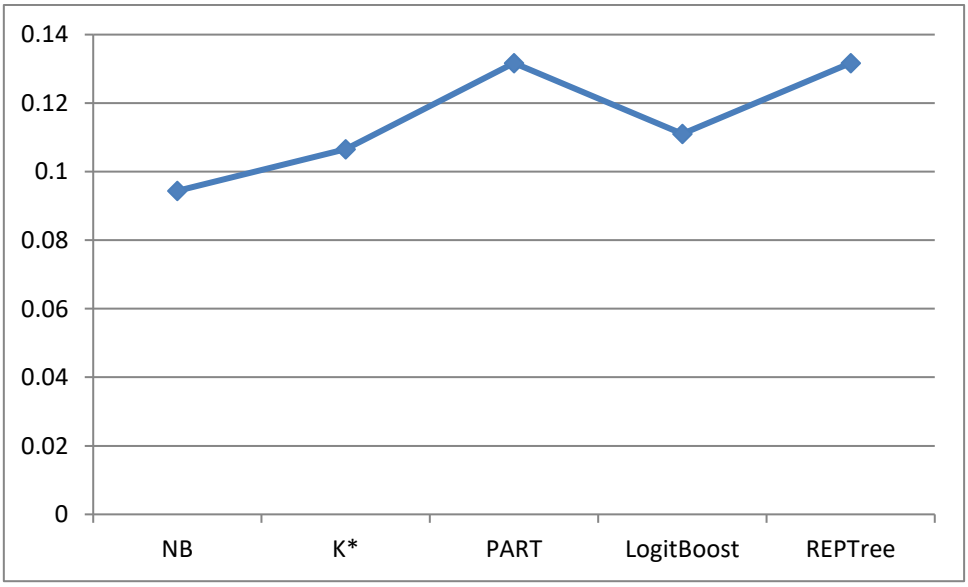

Fig 8. Models MAE for smoking cessation

A comparison of time taken to build models for student smoke cessation presented in fig 9.

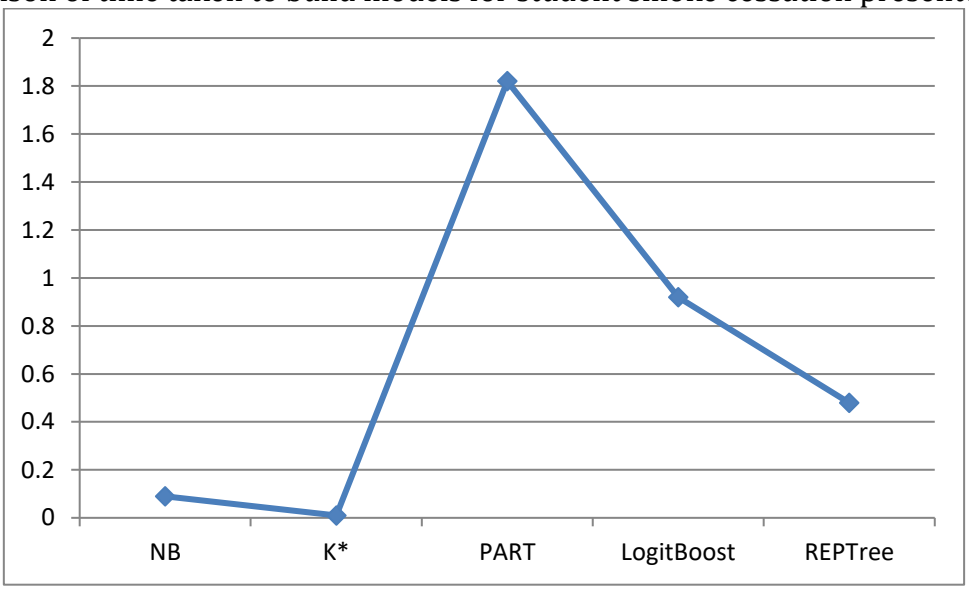

Fig 9. Time taken to build models for smoking cessation (in seconds)

3.3 School environment impact

Students in Iraq spend 4,5 to 5,5 hours in school for 5 days weekly in this study we analyzed the impact of school environment that can lead to nicotine addiction or even trying any tobacco product.

The self-reported questions related to school enrolled in this section to predict if the student trying cigarettes (even one or two buffs), the performance of the five models with school environment impact showed in Table 6.

Table 4. Evaluation of machine learning models for school impact

	AC	KS	MAE
NB	76.9787	0.493	.2566
K*	73.6596	0.3829	0.3075
PART	75.2766	0.4516	0.3045
LogitBoost	78.383	0.05062	0.2949
REPTree	77.5745	0.4919	0.3052

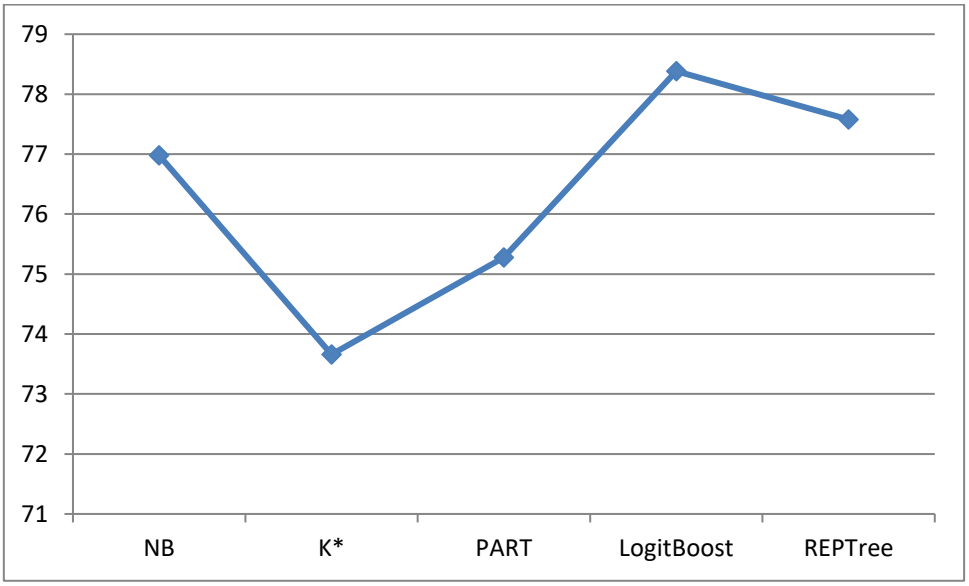

Fig 10. Models accuracy for school impact

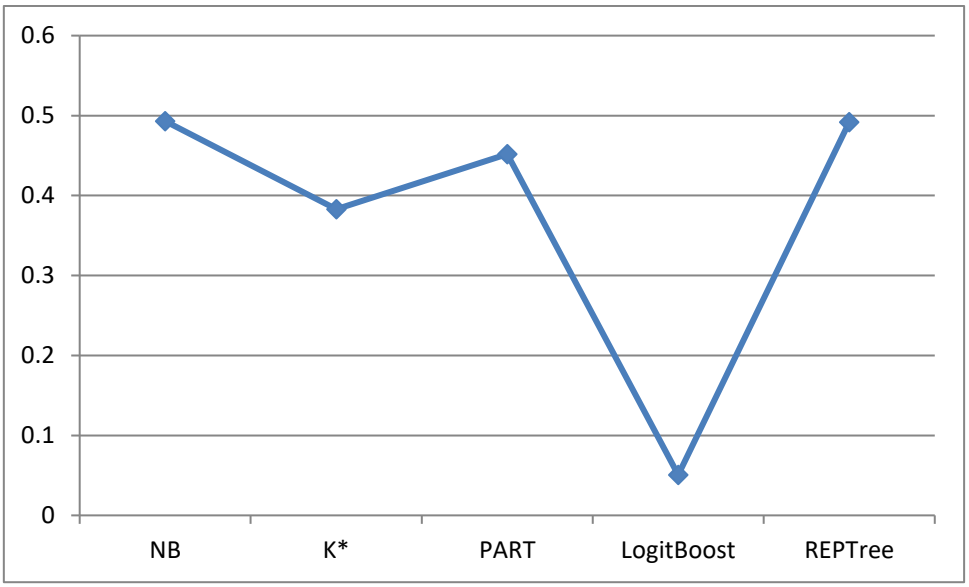

Fig 11. Models kappa statistic for school impact

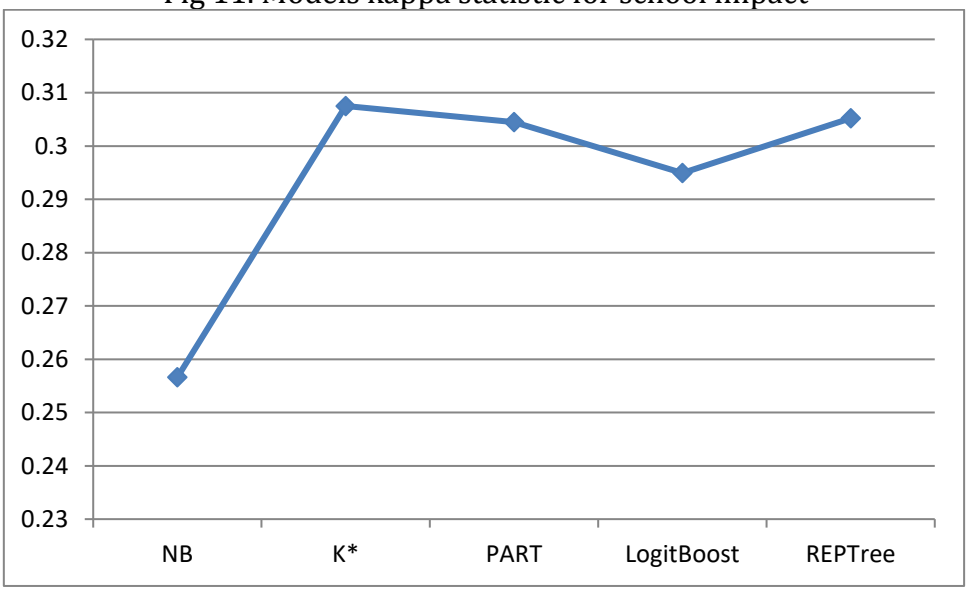

Fig 12. Models MAE for school impact

Fig 13. Shows a comparison of time taken to build models for the impact of school environment that led students to try any tobacco products.

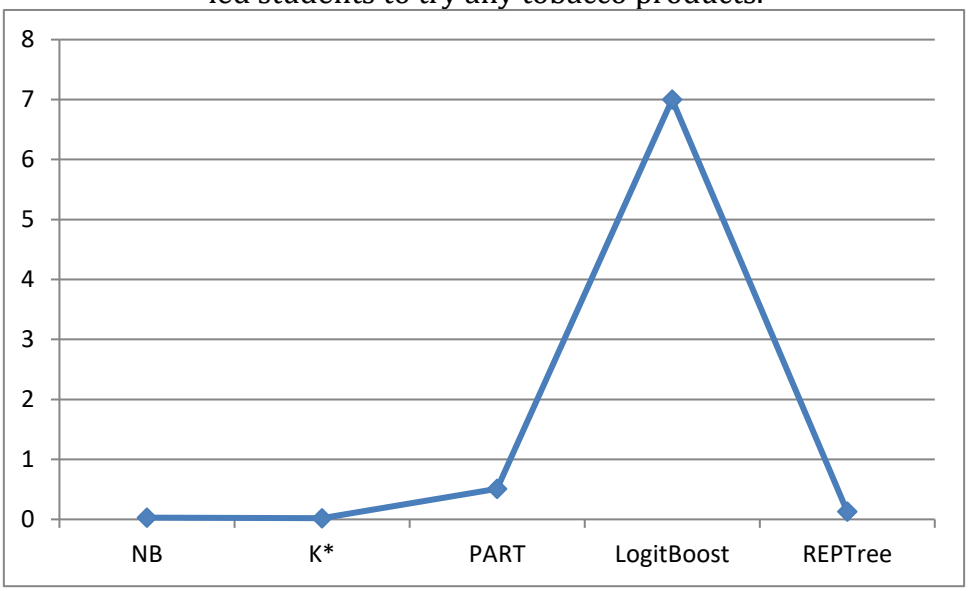

Fig 13. Time taken to build models for school impact (in seconds)

4. Discussion

In this paper supervised machine learning models trained and evaluated to study smoking behavior for a certain age (group of individuals). The data collected from Iraqi school students by answering self-reported questions that may address the reasons led to that behavior in such young age.

The aim of this study is to: (1) predict over all smoking behavior , (2) predict the smoking cessation ability, and (3) analyze the impact of school environment. Five models trained and showed promising outcomes in the process of understanding smoking behavior, LogitBoost scored best accuracy in predicting smoking cessation and in analyzing school environment impact, this achieved through using selected report questions related to each vector. K* model achieved best accuracy in prediction the student behavior through applying all reported questions.

On the other hand for time taken to build each predictor Naïve Bayes take less time to predict the first vector, K* score less time in prediction smoking cessation and school impact.

5. Conclusions

This paper presents a machine learning method to study the smoking behavior of school students in Iraq. LogitBoost showed best performance in predicting student smoking behavior from two perspectives, smoking cessation and the school environment impact. On the other hand K* model achieved best results in predicting overall smoking for students. These results proves that we can analyze and understand this behavior in young individuals to help them quit smoking and use school impact to improve its roll in raising the awareness of smoking hazards.

References

[1] W. H. Ban, C. D. Yeo, S. Han, and S. Kang, "Impact of smoking amount on clinicopathological features and survival in non-small cell lung cancer", Ban et al. BMC Cancer 20:848, 2020.

[2] S. S. Thakur, P. Poddar, and R. B. Ray, "Real-time prediction of smoking activity using machine learning based multi-class classification model", Multimedia Tools and Applications 81:14529–14551, 2022.

[3] Y. Xu, L. Cao, X. Zhao, and L. Li, "Prediction of Smoking Behavior From Single Nucleotide Polymorphisms With Machine Learning Approaches", volume 11, article 416, 2020.

[4] M. M. Hasan, N. Hasan, M. S. Alsubie, and M. M. R. Komol, "Diagnosis of Tobacco Addiction using Medical Signal: An EEG-based Time-Frequency Domain Analysis Using Machine Learning", Advances in Science, Technology

and Engineering Systems Journal Vol. 6, No. 1, 842-849 (2021).
[5] C. Fan, A. F. Gao, "A New Approach for Smoking Event Detection
Using a Variational Autoencoder and Neural Decision Forest", IEEE Access, volume 8, 2020.
[6] Z. Wang, A. Boueie, and P. J. Castaldi" Improved prediction of smoking status via
isoform-aware RNA-seq deep learning models", PLOS computational Biology vol 17(10), 2021.
[7] B. R. Raiff, C. Karatas, and E.A. McClure, " Laboratory Validation of Inertial Body Sensors to Detect Cigarette Smoking Arm Movements", Electronics, ISNN 2079-9292, 3, 87-110, 2014.
[8] M. H. Imtiaz, R. I. Ramos-Garcia, V. Y. Senyurek, and S. Tiffany, "Development of a Multisensory Wearable System for Monitoring Cigarette Smoking Behavior in
Free-Living Conditions", MDPI Electronics, 6, 104, 2017.
[9] P. Mamoshina, K. Kochetov, F. Cortese, and A. Kovalchuck,"Blood Biochemistry Analysis to Detect Smoking Status and Quantify Accelerated Aging in Smokers", DOL: 10.1038/s41598-018-35704, 2019.
[10] P. Lopez-Meyer, S. Tiffany, and E. S. Senior, "Monitoring of Cigarette Smoking Using Wearable Sensors and Support Vector Machines", IEEE, Vol. 60, No. 7, 2013.
[11] A. H. Krist, K. W. Dovidson, and C. M. Mangione, "Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Persons", JAMA, Vol. 325, No. 3, 2021.
[12] N. D. Freedman, D. T. Silverman, and A. R. Hollenbeck, "Association Between Smoking and Risk of Bladder Cancer Among Men and Women", JAMA, Vol. 306, No. 7, 2011.
[13] V. A. Menshov, A. V. Trofimov, and A. V. Zagurskaya, "Influence of Nicotine from Diverse Delivery Tools on the Autonomic Nervous and Hormonal Systems" Biomedicines, 2022.
[14] J. Choi, H. Jung, A. Ferrell, and L. Haddad, " Machine Learning-Based Nicotine Addiction Prediction Models for Youth E-Cigarette and Water pipe (Hookah) Users" Clinical Medicine, 10, 972, 2021.
[15] H.M. and S.M.N,"A Review on Evaluation Metrics for Data Classification Evaluation", IJDKP. Vol. 5, no. 2, 2015.
[16] M. Yusa, " Classifiers evaluation: Comparison of performance classifiers based on tuples amount", IEEE, ISBN:**978-1-5386-0550-9**, 2017.