

Predicting Student Smoking Habits with Machine Learning Techniques

Shashank H M^{1*} and Hemanth Kumar²

MCA Student, Department of MCA¹

Associate Professor, Department of MCA²

Jawaharlal Nehru National College of Engineering, Shimoga, Karnataka, India

shashankholaeyvar@gmail.com and hemanthkumar@jnnce.ac.in

Abstract: *Smoking among students remains a health concern, the intent of this work is to predict the students those who smokes cigarette by utilizing a machine learning-based approach based on behavioral, socioeconomic and other factor. The model should utilize available data to accurately classify students into smoker and non-smoker categories. The data is collected through Google forms from random people. Here in this work Random Forest models, Logistic Regression techniques, and Decision Tree methods are employed for building prediction model for smoking behavior. Comparative analysis of these three algorithms provides us the vision on which method is significant for the task.*

Keywords: Smoking

I. INTRODUCTION

Smoking is considered as one of these harmful impacts on both health and expenses. The well-being of students is still affected by whether they smoke while they're learning. It was also noted that promising Machine Learning (ML) methods are employed in a growing number of fields for the the reason for using predictive modeling in recent years, despite these methods were initially introduced to health sciences. This work predicts smokers among students based on several factors including the following aspects: behavioral, socioeconomic and other features. One of these primary concerns faced in this work is data collection. Collecting accurate data on students' smoking habits, requires honest responses. The information leveraged for this study collected using Google forms, an online survey tool provided by Google Inc. Participants were invited to complete the survey electronically through a link distributed via email and social media platforms. This data is utilized for the purposes of both training and testing models, listed below are three approaches that have been employed: Random Forest models, Logistic Regression techniques, and Decision Tree methods. Comparative analysis algorithms provides the perception which algorithm is significant for the task.

This work helps us to discover what are all the factors that making students to get addicted to smoking habits. With knowledge of these factors help students to stay away and not to fall under these habits and to enhance their academic performance. This project can utilize in

schools and public health programs to find students who might start smoking and help them early. It also helps researchers understand why students smoke and helps create better policies and healthcare practices to reduce smoking.

II. EXISTING SYSTEM

On the concept of "Prediction of smokers among students using machine learning techniques" many work have been completed from different authors Mona Issabakhsh et al. in [1] work by focuses regarding the impact of social factors on the smoking behaviour of college students. It aims to identify key predictors of smoking cessation readiness among this demographic. Carla J. Berg et, al. in [2] projected in examining the characteristics and factors related to smoking frequency among college students and to evaluate the readiness to stop smoking based on different smoking levels. Warren K. Bickel et. al. in [3] analyse quitting smoking process by identifying significant determinants and predicting smoking cessation one year later among participants of the population Assessment of Tobacco and Health (PATH) survey. Liyu Cao et.al. in [4] utilize machine learning techniques, particularly support vector machines (SVM), random forest (RF) methods, to build predictive models for smoking behaviour based on genomic profiles, aiming to predict

individuals' inherited predisposition to smoking by analysing single nucleotide polymorphisms (SNPs). Authors in [5] The objective of this study is to cultivate or create model utilizing the decision tree algorithm in machine learning to predict the daily smoking times of individuals. The study aims to enhance the accuracy of smoking behaviour analysis and identify optimal intervention times, concluding that the XGBoost algorithm achieved optimal performance with a rate of accuracy 84.11%. Authors in [6], Design machine learning models aimed to predict smoking cessation outcomes, using data from 4875 patients in Northern Taiwan, to provide personalized success rates and improve smoking cessation treatments. John Pierce et. al. in [7] predict future smoking behaviour in teenagers, independent of their past smoking experience, suggesting their importance in tobacco control evaluations. Sunday Azagba et. al. in [8] examines the association between age at first puff and age at first whole cigarette and current smoking status among students in Canadian high schools revealing that both measures significantly predict current smoking behaviour, underscoring the significance of preventing early smoking initiation in youth. Work in [9] explores predictors of smoking cessation success among low-income individuals undergoing group CBT, using an approach in machine learning identify key characteristics linked to treatment response, indicating potential for personalized smoking cessation interventions. Charles Frank et. al. in [10] explores Machine learning algorithms employed for predicting smoking behavior based on medical data, revealing significant differences In blood test readings comparing smokers and non-smokers. Logistic regression outperforms other algorithms, offering potential to enhance patient assessment and treatment in healthcare settings. Shreerudra Pratik et. al. in [11] addresses the increasing prevalence of smoking addiction among youths by employing machine learning, deep learning methodologies to predict e-cigarette addiction. Utilizing elastic net regression and K-Nearest Neighbour algorithms, the hybrid the prediction model demonstrates high accuracy in feature selection and prediction, offering insights into demographic factors influencing smoking addiction. Saurabh Singh Thakur et. al. in [12] develops machine learning framework to detect smoking activity in real-time using wrist- wearable sensors. By processing streaming sensor data and employing classification models, predictive accuracy reaches up to 98.7%. The findings facilitate timely interventions for smoking cessation, offering potential applications in preventive healthcare and recognition of other activities of interest.

III. METHODOLOGY

Here for predicting smokers among students, Machine Learning strategies used. A range of machine learning approaches including K-NN, CNN, Logistic Regression, so on. Here in this work Random Forest models, Logistic Regression techniques, and Decision Tree methods are used for prediction.

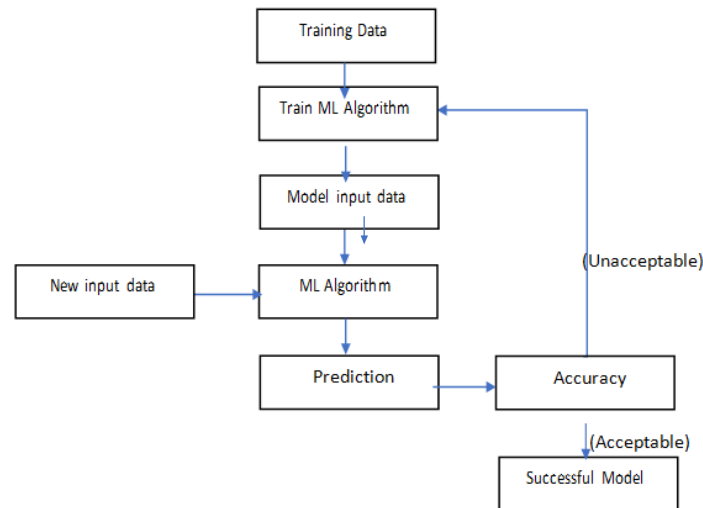


Fig 1: Machine Learning Model

Fig. 1 describes the workflow of ML model.

- **Training Data-** The training data is responsible for the model's accuracy. Training data is equivalent to or more than 60% of the total data.
- **Train ML Algorithm-** By using the training data the ML algorithm is trained. During this time the algorithm identifies the difference in the data
- **Model Input Data-** After training data the input data is fed into the trained model in next step. It is same type of data that is used during training process but not observed by the model during training.
- **ML Algorithm-** In this step the ML algorithm is selected based on our problem statement. Here Random Forest models, Logistic Regression techniques, and Decision Tree methods are used.
- **New Input Data-** After once the algorithm is selected, the upcoming action is to make predictions on the new input data. The new input data is different from the training input.
- **Prediction-** In this step the prediction is done based on new input data. These prediction gives the output of model.
- **Accuracy-** The model's performance is evaluated using a term called Accuracy. Accuracy measures how good the model predicts the output for the given new input data.
- **Successful Model-** If the model is giving good level of accuracy, it is said to serve as a successful model. If the accuracy is not satisfied again the steps are repeated from Training ML algorithm.

Random Forest algorithm

Random Forest a supervised learning Machine Learning algorithm, it can utilized for both classification as well as regression tasks. Random Forest classifier relies on majority voting of decision trees. Decision trees serve as the fundamental components in Random Forest classifier Technique.

Decision tree represents a flowchart-like structure used to make decisions or predictions. Decision Tree basically has 3 nodes

- **Root Node:** Represents the entire dataset and the initial decision to be made.
- **Internal Nodes:** Represent decisions or tests on attributes. Each internal node contains or more branches.
- **Leaf Nodes:** Represent the ultimate decision or prediction. No further splits occur at these nodes

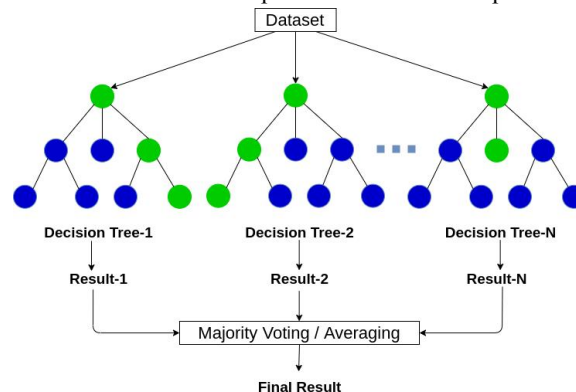


Fig 2: Random Forest Classifier

Fig. 2 explains the Random Forest Algorithm, which starts with a dataset, the assemblage of data used for making predictions. Several decision trees are built using the dataset. Each decision trees undergo trained and every decision trees provides their individual prediction based on available data.

Random Forest classifier combines the result from all the decision trees. In final result, it uses the majority voting prediction among the decision trees for Classification tasks. For regression tasks, it averages the predictions of decision trees as the concluding result.

Proposed System Workflow

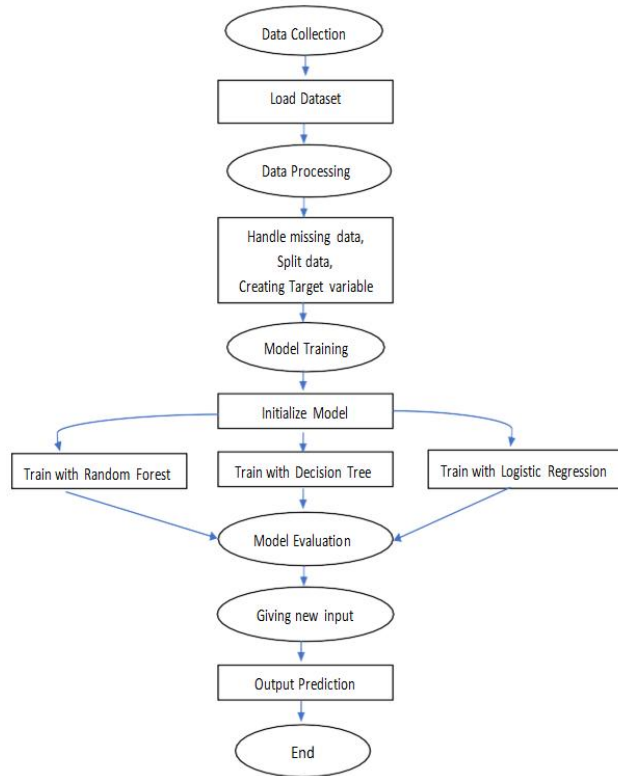


Fig 3: Proposed System Workflow

Fig. 3 illustrates the workflow of the proposed system, which begins with collecting data finishes with prediction-making data collection comes before data processing which manages missing data split the data and creates a target variable once the data is fully processed Random Forest models, Logistic Regression techniques, and Decision Tree methods are employed train the models. Each model is tested so to gauge how successfully it works, the models can generate predictions with new input data after undergoing training and evaluated.

IV. RESULTS

Training and Testing Accuracies of Models

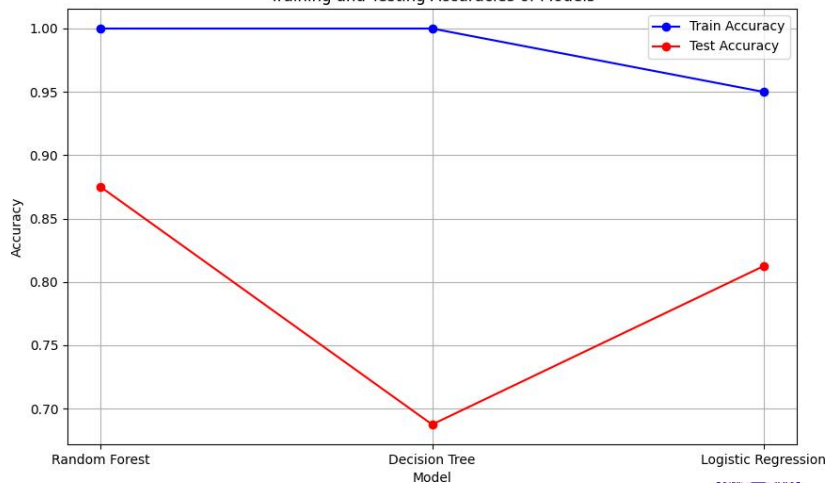


Fig 4: Training and testing accuracy

In the Fig. 4 the training and testing accuracy of logistic regression algorithm is shown. X-axis represents the quantity of trees, and Y-axis Represents the accuracy percentage ranging from 0% to 100%.

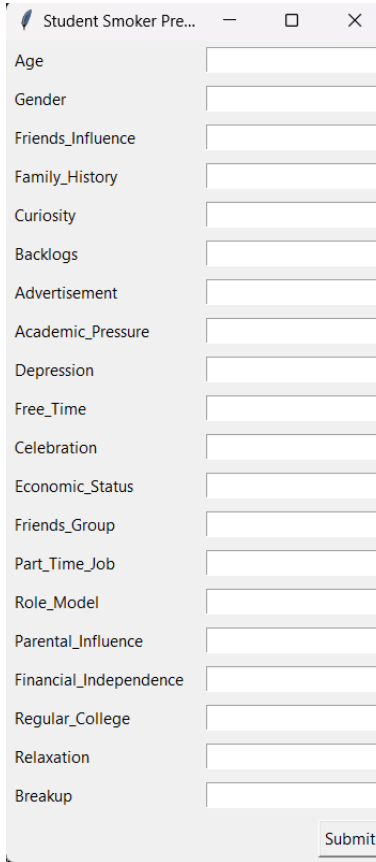


Fig 5: User interface

Fig. 5 is the User interface. After filling all the parameters, we click on the submit button for prediction. Once the submit button is pressed, the model predicts the output as well as provides whether a student is a smoker or not.

V. CONCLUSION

Machine Learning strategies are employed to forecast which students are inclined to smoke cigarettes based on behavior and socioeconomic level among other criteria. Decision Trees Logistic Regression, Random Forest are utilized three distinct algorithms. Upon comparing these algorithms, we discovered that random forest algorithm method is more accurate in predicting students smoking patterns. According to our work, schools and public health initiatives can pinpoint students who are at risk of smoking early stage of habit and offer them support to stay away from smoking.

REFERENCES

- [1] Mona IssabakhshID1, Luz Maria Sa'nchez-Romero1, Thuy T. T. Le2, Alex C. Liber1, Jiale Tan3, Yameng Li1, Rafael Meza4, David Mendez2, David T. Levy1, "Machine learning application for predicting smoking cessation among US adults", 2023, doi: 10.1371/journal.pone.0286883.
- [2] Carla J. Berg, Pamela M. Ling, Rashelle B. Hayes, Erin Berg, Nikki Nollen, Eric Nehl, Won S. Choi, Jasjit S. Ahluwalia, "Smoking frequency among current college student smokers: distinguishing characteristics and factors related to readiness to quit smoking", Health Education Research, Vol.27, Issue 1, 2012, <https://doi.org/10.1093/her/cyr106>.

- [3] Warren K. Bickel, Devin C. Tomlinsona, William H. Craft, Manxiu Maa, Candice L. Dwyer , Yu-Hua Yeha, Allison N.Tegge, Roberta Freitas-Lemos , Liqa N. Athamneha, “Predictors of smoking cessation outcomes identified by machine learning:A systematic review”, Volume 6, 100068, ISSN 2772-3925, 2024, <https://doi.org/10.1016/j.addicn.2023.100068>.
- [4] Yi Xu1, Liyu Cao, Xinyi Zhao, Yinghao Yao, Qiang Liu, Bin Zhang, Yan Wang1, Ying Mao1, Yunlong Ma, Jennie Z. Ma, Thomas J. Payne, Ming D. Li and Lanjuan Li, “Prediction of Smoking Behavior From Single Nucleotide Polymorphisms With Machine Learning Approaches”, *Frontiers in Psychiatry*, Volume 11, ISSN 1664-0640, 2020, doi: 10.3389/fpsyt.2020.00416.
- [5] Yupu Zhang, Jinhai Liu, Zhihang Zhang, Junnan Huang, “Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm”, 2019.
- [6] Cheng-Chien Lai, Wei-Hsin Huang, Betty Chia-Chen Chang and Lee-Ching Hwang, “Development of Machine Learning Models for Prediction of Smoking Cessation Outcome”, *International Journal of Environmental Research and Public Health*, 2021, <https://doi.org/10.3390/ijerph18052584>.
- [7] Melanie Wakefield, Deborah D. Kloska, Patrick M. O’Malley, Lloyd D. Johnston, Frank Chaloupka, John Pierce, Gary Giovino, Erin Ruel & Brian R. Flay, “The role of smoking intentions in predicting future smoking among youth: findings from Monitoring the Future data”, 2003, doi:10.1111/j.1360-0443.2004.00742.x.
- [8] Sunday Azagba, Neill Bruce Baskerville, Leia Minaker, “A comparison of adolescent smoking initiation measures on predicting future smoking behavior”, Volume 2, ISSN 2211-3355, 2015, <https://doi.org/10.1016/j.pmedr.2015.02.015>.
- [9] Lara N. Coughlin, Allison N. Tegge, Christine E. Sheffer & Warren K. Bickel, “A machine-learning approach to predicting smoking cessation treatment outcomes”, *Nicotine & Tobacco Research*, Volume 22, Issue 3, 2020, <https://doi.org/10.1093/ntr/nty259>.
- [10] Charles Frank, Asmail Habach, Raed Seetan, Abdullah Wahbeh, “Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis”, Volume 3, Issue 2, 2018, DOI: 10.25046/aj030221.
- [11] Shreerudra Pratik, Debasish Swapnesh Kumar Nayak, Rajendra Prasath and Tripti Swarnkar, “Prediction of Smoking Addiction Among Youths Using Elastic Net and KNN: A Machine Learning Approach”, 2022, DOI: 10.1007/978-3-031-21517-9_20.
- [12] Saurabh Singh Thakur, Pradeep Poddar & Ram Babu Roy, “Real-time prediction of smoking activity using machine learning based multi-class classification model”, Volume 81, 2022, <https://doi.org/10.1007/s11042-022-12349-6>