IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

Machine Learning Classifies the Smoking and Drinking Impact

Kottur Siddaraju Gari Sai Kiran[1] and Prof. Mrs. Usha Sree R[2]
Student MCA, IVth Semester[1]
Associate Professor, Department of MCA[2]
Dayananda Sagar Academy of Technology and Management, Udaypura, Bangalore, Karnataka, India
saikirankottur7@gmail.com

Abstract: *This research main objective of this paper is to show the impact of the smoking and drinking habits on our body and making comparative analysis between algorithms to find out which algorithm provides better accuracy. With the help of techniques such as EDA we find out the relationship between the smoking and drinking habits and how interrelated one habit is to another and with the help of Data Visualization techniques we'll be using such as graphs and other methods to show data visually for better understanding. Algorithms like Logistic Regression, Random Forest, XGBoost, LGBM classifier are used for comparative analysis and the main intention of the paper is to raise awareness in the society that how dangerous these habits are, showing them the impact of these habits having on our body.*

Keywords: Addiction, Diseases, Drinking, Machine Learning, Smoking, Smoking and Drinking

I. INTRODUCTION

The lifestyle of humans has been changing constantly as the time goes forward, as technology is growing it's affecting the lifestyle of humans in one way or other. Such as bicycle to cars, immovable telephones to portable smartphones, from smoking tobacco to smoking weed. But the thing that's been constant in our lifestyles is these habits of smoking and drinking. They got involved way too much into our lives that they became basic necessities such as food, water, and shelter. Even though there's been so many policies and awareness programs to stop doing them, the addiction is being too hard to break those habits. These habits are surely one of the reasons if any of the individual is suffering from any kind of disease. The scope of the diseases that an individual can get through these habits is too large that it's getting hard to comprehend just how much these habits are affecting our body. So the ultimate objective of this paper is to use Machine Learning techniques to show impact of these habits having on our body, correlation between these habits and making comparative analysis between algorithms to check which gives better accuracy, through which we can treat further better for the individuals that are suffering from diseases.

PROBLEM STATEMENT

It is difficult to forecast the combined effects of smoking and drinking on individuals. However, traditional analytical approaches are insufficient in mining complex patterns from huge datasets thus undermining interventions and public health strategies. This study seeks to use machine learning models to analyze and predict the impacts of smoking and drinking on an individual's health status. We hope that by evaluating different algorithms that can be used within machine learning environments for this purpose, we may find the best models for predicting health outcomes with some policy implications in terms of improving people's lives.

II. LITERATURE SURVEY

Since smoking is the leading preventable cause of death and disease in the industrialized world, developments in machine learning and modern electronics may enable us to provide smokers with innovative, real-time interventions. In this work, we examine different machine learning methods to efficiently define high-urge states using situational parameters associated with the presence or absence of smoking impulses during a cessation attempt. In order to evaluate our machine learning techniques—Bayes, discriminant analysis, and decision tree learning methods—we employed a

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

dataset gathered from more than 300 individuals who had made a quit attempt. The evaluation of the three categorization strategies takes into account factors including sensitivity, specificity, accuracy, and precision. The analysis's conclusion demonstrated that feature selection-based algorithms can achieve excellent classification rates even when only a small number of features are chosen from the complete dataset. With an 86% classification accuracy, the classification tree technique performed better than the naive Bayes and discriminant analysis methods. These data point to a potential application for mobile health applications: machine learning may be a useful strategy for addressing issues related to smoking cessation and for anticipating the impulse to smoke. In conclusion, the search for the optimum features and classifier parameters greatly enhances the algorithms' performance, and machine learning classifiers can assist in identifying smoking scenarios. Furthermore, this research demonstrates the value of new technology in enhancing the efficacy of smoking cessation programs, therapists' time and patient management, and ultimately the efficiency of the health care system. Future research should concentrate on creating innovative expert systems that can give real-time treatments, with the goal of giving more individualized and adaptive help to those who truly need it iin the quickest possible time [1].

The improvement of the algorithm and the accumulation of smoking data have made exact analysis possible, which has significant benefits for smoking cessation. But to the best of our knowledge, not much research has been done on the habits of regular smokers, such as the exact moment they light up. This paper proposes a model based on a decision tree machine learning method to anticipate daily smoking time. The Chinese Center for Disease Control and Prevention's population data on smokers was used to create the simulation data set of smoking time data. To tackle the problem of inadequate feature information, we offer an extraction module for feature information. We concluded that, with an accuracy rate of 84.11%, the XGBoost-based prediction model performed best and could be trained significantly more quickly than models based on other machine learning algorithms.[2]

Smoking and alcohol usage are two of the most preventable causes of death in the globe. Early detection can be extremely important in the fight against these forms of addictions. Our study uses bodily signals to create a machine-learning-based classification system that predicts the types of drinkers and smokers. Naïve Bayes, Random Forest, K-Nearest Neighbors, and Logistic Regression were utilized for the classification. We have noticed that out of the four classification algorithms we have used, Naïve Bayes and Random Forest both perform well. In this sense, Random Forest outperforms Naïve Bayes. Random Forest is the best model out of the four in terms of accuracy, precision, recall, and f1 score, scoring the highest for both drinker and smoker categories Rand. Taking everything into account, we may conclude that Random Forest would be an excellent option for forecasting the kinds of people who drink and smoke.[3]

First of all, Although it is common in the US, unhealthy drinking is not well recognized or addressed. For primary care physicians, identifying individuals who are unhealthy drinkers can be uncomfortable and time-consuming. By concentrating attention on patients who are most likely to require care, an automated identification rule could improve efficiency and efficacyThe purpose of this study was to use widely available laboratory and demographic data to create a clinical prediction tool for binge drinking. We collected 38 laboratory and demographic variables on 43,545 nationally representative people who had available alcohol consumption data as a reference standard from the National Health and Nutrition Examination Survey (1999 to 2016). Clinical prediction models were constructed using neural networks, decision trees, random forests, k-nearest neighbor, logistic regression, and support vector machines. In order to construct the prediction tool, the model with the largest area under the receiver operator curve was chosen. In the test set, the random forest model with 15 variables generated the biggest area under the receiver operator curve (0.78). High-density lipoprotein, age, hemoglobin, sex, and current smoking were the most significant predictors.It was discovered that the optimal operating point had the following values: 0.50 for sensitivity, 0.86 for specificity, 0.55 for positive predictive value, and 0.83 for negative predictive value. The tool's application led to a 75% reduction in the target sample. A decision tool can potentially increase the efficiency and reach of screening by identifying a group of individuals who appear to need clinical care for unhealthy drinking based on data that is widely available.[4]

Adolescent alcohol use is a complicated experience, with regional variations in the age at which alcohol is purchased as well as individual variations in use and outcomes. The significance of predicting teenage alcohol consumption through modeling is highlighted by the latter. To predict different alcohol consumption levels in mid-adolescence, the current study (a) evaluated the effectiveness of seven machine learning methods. (b) Using a cross-cultural cross-study scheme

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

in the training-validation-test process, the best performing machine learning algorithm's predictive power was demonstrated. Seven machine-learning methods are compared: neural networks, logistic regression, support vector machines, random forests, lasso regression, ridge regression, and elastic-net. The algorithms employed a number of prediction indices, including area under the curve (AUC), recall, accuracy, precision, negative predictive value, and F1 prediction score. The elastic-net machine-learning system demonstrated the best predictive performance in both the Australian (AUC = 0.855 ± 0.072) and Canadian (AUC = 0.869 ± 0.066) samples, according to prediction indices. The domain contribution analysis showed that the models with only personality clusters (AUC = 0.776 ± 0.063/0.796) and only psychopathology (AUC = 0.816 ± 0.044/0.790 ± 0.071 ± 0.066 in Canada/Australia) had the greatest prediction accuracy indices.The models using only psychopathology (AUC = 0.816 ± 0.044/0.790 ± 0.071 in Canada/Australia) and only personality clusters (AUC = 0.776 ± 0.063/0.796 ± 0.066 in Canada/Australia) The domain contribution study revealed that the countries with the greatest prediction accuracy indices were Canada and Australia. Similar to this, externalizing psychopathologies, baseline alcohol use, and the sensation-seeking personality profile all contributed to the prediction in both populations, regardless of the amount of alcohol consumption. The potential of computerized screening technologies to forecast teenage alcohol consumption is evident.[5]

Previous studies have shown that doctors and medical students consume more alcohol. Building machine learning (ML) models to detect patterns of high-risk drinking (HRD), including alcohol use disorder, in this group is the goal of the current study. We looked over data collected through an online poll from medical students in Brazil. The variables included mental health, university status, personal information, and sociodemographic data. The AUDIT-C scores served as the basis for HRD stratification. Elastic net regularization, random forest, and artificial neural networks are the three machine learning algorithms that were utilized to create classifiers to predict HRD in medical students. In order to evaluate the most significant predictors for the decisions made by the models—which could be factors related to HRD—model interpretation approaches were used. The study comprised 4840 medical students in total. A 53.03% prevalence of HRD was observed. The three machine learning models that were developed performed remarkably well in differentiating between people who had high-risk drinking (HRD) and those who did not. The test set exhibited a similar level of performance to the cross-validation technique, with average AUC scores of approximately 0.72. The usage of cannabis and tobacco, monthly family income, marital status, sexual orientation, and physical activity were the most significant characteristics for the ML models. According to this study, machine learning models could be used to assess students at first for alcohol use disorders or at-risk drinking. Furthermore, we have discovered a number of important variables linked to HRD that warrant additional research and consideration for intervention and support strategies.[6]

Grown-ups encountering vagrancy are bound to have a liquor use jumble contrasted with grown- ups in everybody. Treatments based on shelters are common, but few people finish them, indicating that this population, which is understudied and underserved, requires better methods that are tailored to it. One obstruction to growing more powerful medicines is the restricted information on the triggers of liquor use among destitute grown-ups. This paper depicts the utilization of biological flashing evaluation (EMA) to distinguish indicators of "unavoidable drinking" (i.e., drinking inside the following 4 h), among an example of grown-ups encountering vagrancy and getting wellbeing administrations at a destitute sanctuary. Over the course of four weeks, 78 persons, the majority of whom were male (84.6 percent), who were homeless and reported unsafe drinking, completed 4557 EMAs. Their average age was 46.6, and they completed up to five EMAs every day. The study developed a drinking risk algorithm using machine learning techniques that correctly predicted 82% of imminent drinking episodes within 4 hours of the first drink of the day and 76% of nondrinking episodes. The calculation incorporated the accompanying 7 indicators of approaching drinking: inclination to drink, having liquor effectively accessible, feeling sure that liquor would further develop mind-set, feeling discouraged, lower obligation to being without liquor, not connecting with somebody drinking liquor, and being inside. The examination group utilized the outcomes to foster mediation content (e.g., brief custom fitted messages) that will be conveyed while approaching drinking is distinguished in an impending mediation stage. To support positive change, we specifically developed three theoretically grounded message tracks that are further tailored to a participant's current drinking goal (i.e., stay sober, drink less, or no objective) and focus on mood/negative affect, social/availability, and craving/urge. This is the first study to our knowledge to develop personalized intervention messages for adults experiencing homelessness based on likelihood of imminent drinking, current drinking triggers, and drinking goals.[7]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

Noncommunicable diseases caused by smoking (SiNCDs) are now a significant threat to public health and a leading cause of death worldwide. Numerous studies using artificial intelligence to predict the risk of developing SiNCDs have been proposed over the past ten years. In such systems, however, it can be quite difficult to identify the most important characteristics and create models that can be understood.Here, we provide an efficient methodology based on extreme gradient boosting (XGBoost) and population-level SiNCD prediction in South Korea and the US the hybrid feature selection (HFS) technique. In the beginning, HFS is carried out in three stages: (I) significant features are selected using the t-test and chi-square test; (II) dissimilar features are obtained using multicollinearity analysis;and (III) the best representative features are chosen using the least absolute shrinkage and selection operator (LASSO). The XGBoost predictive model is then fed specific features. Our proposed model outperforms several existing baseline models, as demonstrated by the experimental results. The proposed model also includes important features that make the SiNCDs prediction model easier to understand. Consequently, it is expected that the framework based on XGBoost will help with early detection and prevention of SiNCDs in public health concerns.[8]

Smoking and unreasonable liquor utilization stay squeezing general medical problems. The goal of this study was to create machine learning (ML) models that would use sensor and biomarker data to predict drinking and smoking habits in order to target interventions. The 10 Machine Learning (ML) algorithms that are validated for predicting smoking and drinking are trained on the Smoking and Drinking dataset, which was obtained from Kaggle and compiled from self- reported lifestyle information, blood tests, and readings from wearable sensors. The XGBoost model had a notable accuracy of 73.96 percent when predicting drinking status, while the Random Forest model outperformed the others with an accuracy of 79.65 percent when predicting smoking. In general, the ML models exhibit strong capabilities for predicting smoking and drinking based on biological data in real time. Such prescient investigation hold guarantee for early gamble discovery, customized care, and nuanced intercessions by medical services frameworks. This study looked into how ML algorithms could find patterns in lifestyle behaviors and the health outcomes that go along with them. This work also opens up the Blackbox character of ML models through the use of Explainable AI tools, namely "SHAP (SHapley Additive exPlanations)" and "LIME (Local interpretable model-agnostic explanations)."[9]

Premature death is significantly influenced by smoking, which is one of the most significant public health issues. In recent years, a large number of machine learning-based decision support systems have been created to assist people quitting smoking. However, implementing such systems is thought to present a significant obstacle due to the inevitable class imbalance. In order to address the issue of class imbalance in the prediction of smoking cessation interventions among the Korean population, we conduct an empirical comparison of machine learning methods in this paper. The goal of this paper is to use synthetic oversampling techniques, such as the adaptive synthetic (ADASYN) and the synthetic minority over-sampling technique (SMOTE), to improve prediction performance for the class imbalance problem. The experimental design, which has three parts, was able to accomplish this. The lasso method and multicollinearity analysis are used in two phases to select the best representative features. Second, make use of the SMOTE and ADASYN methods to generate the newly balanced data. Thirdly, the gender- and subject-specific prediction models are built using machine learning classifiers. The f-score, type I error, type II error, balanced accuracy, and geometric mean indices are used to demonstrate the accuracy of the prediction models. When SMOTE and ADASYN were utilized, a comprehensive analysis reveals that the best performances were achieved by gradient boosting trees (GBT), random forests (RF), and multilayer perceptron neural network (MLP) classifiers across all subjects and genders. The SMOTE with GBT and RF models' feature importance ratings enhance the interpretability of the decision-support system. In addition, it has been demonstrated that, when it comes to smoking cessation prediction, the presented synthetic oversampling methods paired with machine learning models performed better than baseline models.[10]

III. METHODOLOGY

Data Collection: This projects collected data from various sources like google forms, other research papers, from Hospitals and other medical fields, and authorized data sets.

Data Preprocessing: The collected data will undergo through the procedure of preprocessing such as removing null values, detecting outliers and cleaning the data using libraries such as pandas, numpy and transforms the dataset in a way that will be easy to analyze handling missing values.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

Exploratory Data Analysis (EDA): The dataset will be subjected to EDA following preparation. This EDA process will help to understand the data easily, and makes it easier to identify the patterns between multiple variables. Data Visualization libraries such as matplotlib provides us various functions such as histograms, heat-maps, pie- charts through which our data can be turned into these graphs for better understanding. In our case we'll be using data visualization and EDA to show the impact of the smoking and drinking habits on our body.

Modeling: The main intention of this paper is to compare and check which algorithm provides higher accuracy, so here we are using some well known algorithms such as logistic regression, random forest, XGB classifier and LGBM classifier.

Model Evaluation: The performance of a model is evaluated using these metrics

1. Accuracy: The percentage of correctly classified instances.
2. Precision: The proportion of correctly anticipated positives to all positively projected predictions.
3. Recall: The proportion of all actual positives to true positive predictions.
4. F1-Score: The precision and recall harmonic mean.

ROC-AUC: The ROC AUC is a measure of how well a model distinguishes between classes.At different threshold values, it shows the true positive rate (TPR) against the false positive rate (FPR). When the AUC is 0.5, there is no discriminative ability, whereas a value of 1.0 indicates perfect discrimination.

Logistic Regression: Logistic regression is a statistical method used for binary classification that models the probability of a binary outcome based on one or more predictor variables. It uses the logistic function to squeeze the output between 0 and 1, making it suitable for predicting probabilities.

Random Forest(RF): Using random selections of training data and features, Random Forest is an ensemble machine learning technique that mixes numerous decision trees to increase predicted accuracy and decrease overfitting.

XGBoost (Extreme Gradient Boosting): XGBoost, often known as Extreme Gradient Boosting, is a fast and effective gradient boosting method implementation. It builds an ensemble of trees sequentially, where each tree corrects the errors of its predecessors

LGBM Classifier (LightGBM): LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be highly efficient and scalable, handling large datasets with lower memory usage and faster training speed compared to traditional gradient boosting methods.

Model	Accuracy	Precision	ROC-AUC	Recall	F1 Score
Logistic Regression	0.70	0.70	0.70	0.71	0.70
Random forest	0.72	0.71	0.72	0.73	0.72
XGBoost	0.71	0.73	0.71	0.73	0.72
LGBM Classifier	0.71	0.70	0.71	0.73	0.72

T a b l e 1 . Accuracy values of models before boosting them .

We got the best results with RF, LGBM and XGB models. We can tune their hyperparameters to get better results.

After Tuning Hyperparameters:

Model	Accuracy
Random Forest	0.71
XGBoost	0.72
LGBMClassifier	0.72

Table 2. Accuracy values of models after boosting them

IV. RESULTS AND DISCUSSIONS

We're able to find the relation between smoking and drinking habits and how greatly it impacts on our body. Through the help of data visualizations and other methods this paper demonstrates how much impact does these habits have on our body and able to make comparative analysis like out of which algorithm and frameworks are more suited to make this type of analysis.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

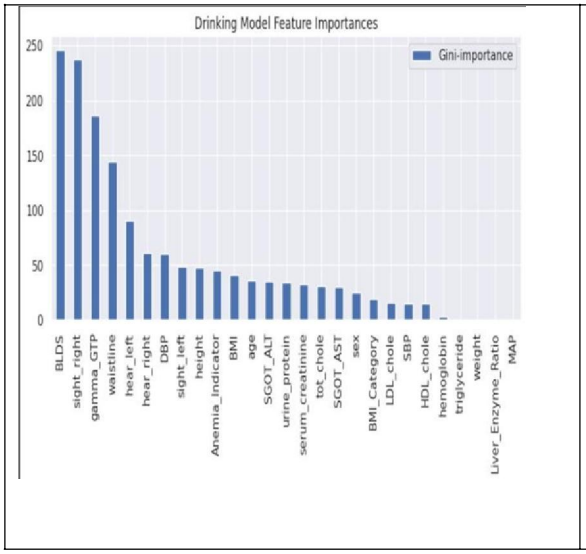

Figure 1. Impact of Alcohol Drinking on our body

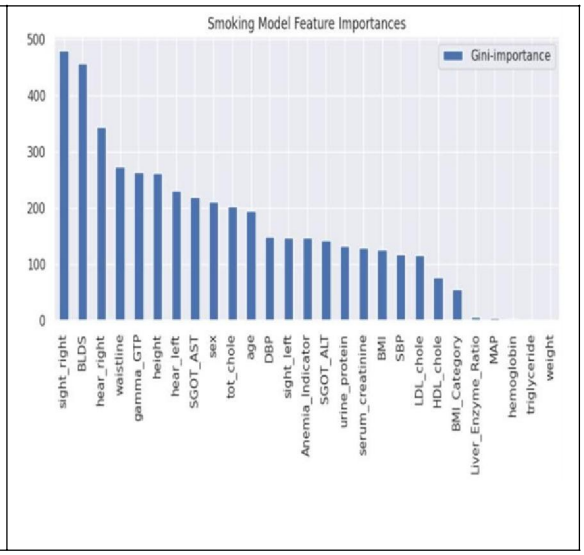

Figure 2. Impact of smoking on our body

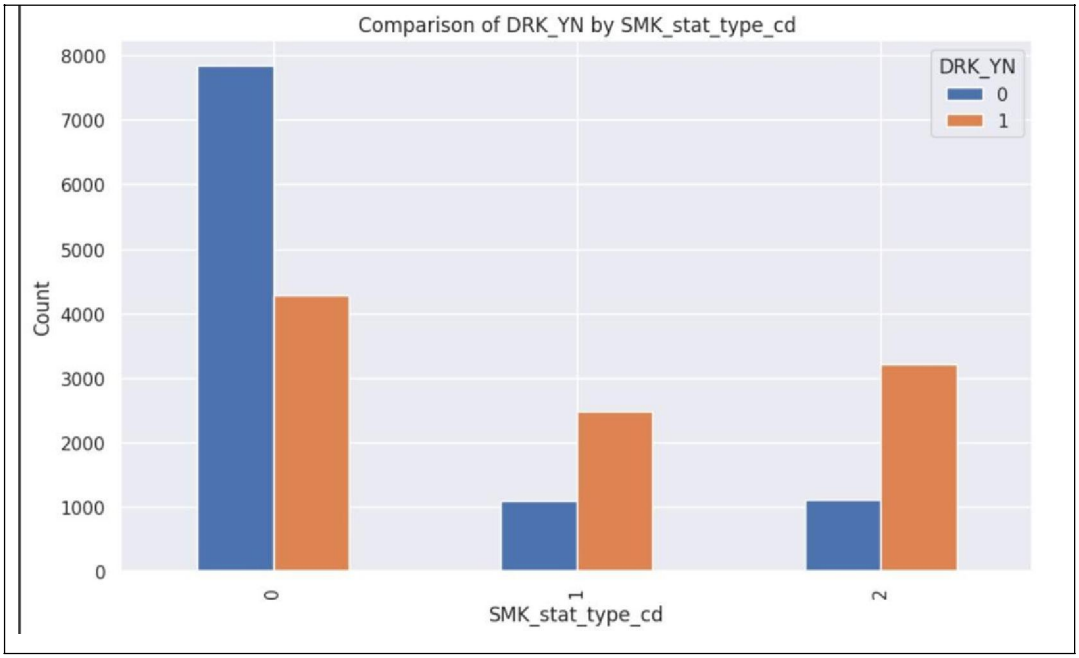

Figure 3. The interrelation between smoking and drinking habits

The LGBM classifier model shown more promising in nature by providing the ROC AUC of 72% for drinking and for smoking with ROC AUC value of 84% making it by far the best model. So instead of using always popular algorithms such as logistic regression, random forest we should be using algorithms that are suitable for our needs. We could further increase the accuracy by using other various boosting methods but this paper main intention is just to do comparative analysis between algorithms but not increasing accuracy so we'll stop here.

V. CONCLUSION

This paper demonstrates the effects of smoking and drinking habits on human body and the interelation between those habits. We also used different algorithm models to check which one is best suitable to do the analysis.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, November 2024

REFERENCES

[1] Antoine Dumortier, Ellen Beckjord, "Classifying smoking urges via machine learning"Computer Methods and Programs in Biomedicine Volume 137, December 2016, Pages 203-213
https://www.sciencedirect.com/science/article/abs/pii/S0169260716304904
[2] Yupu Zhang; Jinhai Liu; et.Al; "Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm" https://ieeexplore.ieee.org/abstract/document/8784698
[3] Rakib Hasan; Ferdous Hasan;"Classifying Different Types of Smokers and Drinkers by Analyzing Body Signals using Machine Learning" The Journal of the American Board of Family Medicine May 2020, 33 (3) 397-406
https://ieeexplore.ieee.org/document/10499139
[4] Levi N. Bonnell, Benjamin Littenberg,"A Machine Learning Approach to Identification of Unhealthy Drinking"The Journal of the American Board of Family Medicine May 2020, 33 (3) 397-406
https://www.jabfm.org/content/33/3/397
[5] Mohammad H. Afzali, Matthew Sunderland et.Al "Machine-learning prediction of adolescent alcohol use: a cross-study, cross-cultural validation"Addiction: Volume 114, Issue 4 Pages: 585-768
https://onlinelibrary.wiley.com/doi/10.1111/add.14504
[6] Grasiela Marcon, Flávia de Ávila Pereira et.Al "Patterns of high-risk drinking among medical students: A web-based survey with machine learning" Computers in Biology and Medicine Volume 136, September 2021, 104747
https://www.sciencedirect.com/science/article/abs/pii/S0010482521005412
[7] Scott T. Walters, Michael S. Businelle et.Al "Using machine learning to identify predictors of imminent drinking and create tailored messages for at-risk drinkers experiencing homelessness" Journal of Substance Abuse Treatment Volume 127, August 2021, 108417
https://www.sciencedirect.com/science/article/pii/S0740547221001434
[8] Khishigsuren Davagdorj,Van Huy Pham et.Al "XGBoost-Based Framework for Smoking-Induced Noncommunicable Disease Prediction" https://www.mdpi.com/1660-4601/17/18/6513
[9] Arnav Thakur; Arunbalaji C G, "Interpretable Predictive Modeling for Smoking and Drinking Behavior using SHAP and LIME" https://ieeexplore.ieee.org/abstract/document/10581153
[10] Jong Seol Lee ,Van Huy Pham, "A Comparative Analysis of Machine Learning Methods for Class Imbalance in a Smoking Cessation Intervention" https://www.mdpi.com/2076-3417/10/9/3307