DATA MINING PROJECT

PREDICTING SMOKING AND DRINKING BEHAVIOR USING BODY SIGNAL DATA

PREDICTING SMOKING AND DRINKING BEHAVIOR USING BODY SIGNAL DATA

-Aishwarya Peri

ABSTRACT

Smoking and drinking behaviors significantly impact public health, contributing to preventable diseases and mortality worldwide. This study aims to develop a robust machine-learning-based predictive model using physiological, demographic, and behavioral data to determine smoking and drinking behaviors. The analysis encompasses data preprocessing, feature selection, and model evaluation. Key features influencing drinking behavior include sex, age, gamma_GTP, hemoglobin, and HDL_cholesterol, while smoking behavior is shaped by hemoglobin, gamma_GTP, serum_creatinine, age, and waistline. Decision Trees were employed for feature selection, ranking variables based on their predictive importance. Models such as Logistic Regression, XGBoost, Decision Trees, and MARS were trained and evaluated. XGBoost demonstrated the highest accuracy of 70%, followed by Logistic Regression at 68%, with weighted F1 scores of 0.70 and 0.68, respectively. These findings emphasize XGBoost as the best-performing model for behavioral classification tasks. The developed models were integrated with a GUI for real-time predictions, enhancing their accessibility and applicability.

Section 1: Introduction

Smoking and drinking behaviors are significant contributors to public health concerns, driving preventable diseases and mortality globally. These behaviors are linked to various adverse health outcomes, including cardiovascular disease, liver dysfunction, respiratory disorders, and multiple forms of cancer. The economic burden of treating these conditions and the societal cost of reduced productivity underscore the need for innovative, data-driven approaches to address these behaviors effectively. Predictive modeling using machine learning offers a promising avenue to identify and mitigate these behaviors by leveraging complex datasets to reveal actionable insights.

This study employs machine learning techniques to predict smoking and drinking behaviors based on physiological, demographic, and behavioral data. By integrating advanced predictive algorithms, the project aims to contribute to public health initiatives by identifying individuals at risk and supporting targeted interventions. The dataset used in this study includes crucial physiological markers such as gamma_GTP, hemoglobin, and serum_creatinine, alongside demographic features like age, sex, and waistline. These variables serve as critical predictors for binary (drinking status) and (smoking status) classification tasks.

The methodology is structured across multiple stages. Data preprocessing ensures the dataset is clean, standardized, and suitable for analysis. Feature selection is performed using Decision Trees, which rank features based on their importance for predictive modeling. Machine learning models, including Logistic Regression [6], Decision Trees [7], XGBoost [9], and MARS [8], are evaluated to identify the best-performing algorithms.

XGBoost emerges as the top model, achieving an accuracy of 70% and Logistic Regression follows with 68%. These results highlight the effectiveness of using physiological and demographic data to predict lifestyle behaviors.

A key feature of this study is the development of a simple and user-friendly graphical user interface (GUI).The system allows healthcare associates to input a user's demographic and physiological details, including variables such as age, waistline, cholesterol levels, and blood pressure, into a user-friendly graphical interface. The backend predictive model processes this information to provide two key outputs: drinking status (drinker or non-drinker) and smoking status (Smoker or Non Smoker). The predictions are presented clearly and comprehensively, ensuring accessibility for healthcare professionals to make informed decisions.

The remainder of this paper is structured as follows: Section 2 discusses related work and the used dataset. Section 3 presents proposed work, results and further recommendations. Section 4 concludes the study.

Section 2: Related Work

Significant research has been conducted to predict smoking and drinking behaviors using machine learning and statistical techniques. However, much of the existing work focuses on theoretical frameworks and limited datasets, leaving room for improvement in real-time applicability and user engagement.

Azagba et al. [1] explored the impact of early smoking initiation on future smoking behavior, utilizing sociodemographic data to predict smoking prevalence. Their study highlighted the role of demographic variables in identifying smoking patterns but lacked the inclusion of physiological metrics such as blood pressure and cholesterol levels, which could enhance prediction accuracy.

Similarly, Zhao et al. [2] analyzed alcohol consumption patterns using traditional regression methods. Although their work identified significant predictors, such as liver enzyme levels and cholesterol, it relied on static datasets and did not incorporate advanced machine learning algorithms or real-time predictive capabilities.

Frank et al. [5] employed machine learning models, including Logistic Regression and Decision Trees, to classify smoking status using blood test results and vital signs. The study demonstrated the effectiveness of computational methods, achieving a classification accuracy of 83.44%. Despite these results, the focus was solely on smoking behavior, with no exploration of drinking habits or interactive systems for real-time predictions.

While these studies provide valuable insights, they share common limitations, such as the absence of real-time feedback mechanisms, narrow feature sets, and a lack of user-friendly applications. Most existing research concentrates on isolated behaviors, such as smoking, without addressing the broader interrelation between smoking and drinking behaviors.

Our project addresses these gaps by introducing a comprehensive, user-friendly solution that integrates advanced machine learning algorithms, such as XGBoost [9] and Logistic Regression [6], with extensive physiological and demographic data inputs. Unlike prior work, our application allows users to input data in real time and obtain actionable predictions, bridging the gap between theoretical frameworks and practical usability. Additionally, we gather a wide range of physiological measurements, including cholesterol levels, blood pressure, and liver enzyme metrics, offering a more detailed health analysis compared to studies with limited datasets. The results are presented in a clear and simple format, such as “Drinker” or “Non-Drinker,” and “Smoker” or “Non-Smoker” enabling immediate comprehension and decision-making for healthcare professionals and policymakers. This approach enhances the relevance and practical applications of predictive modeling in healthcare and insurance industries, contributing to more effective risk assessments and public health interventions.

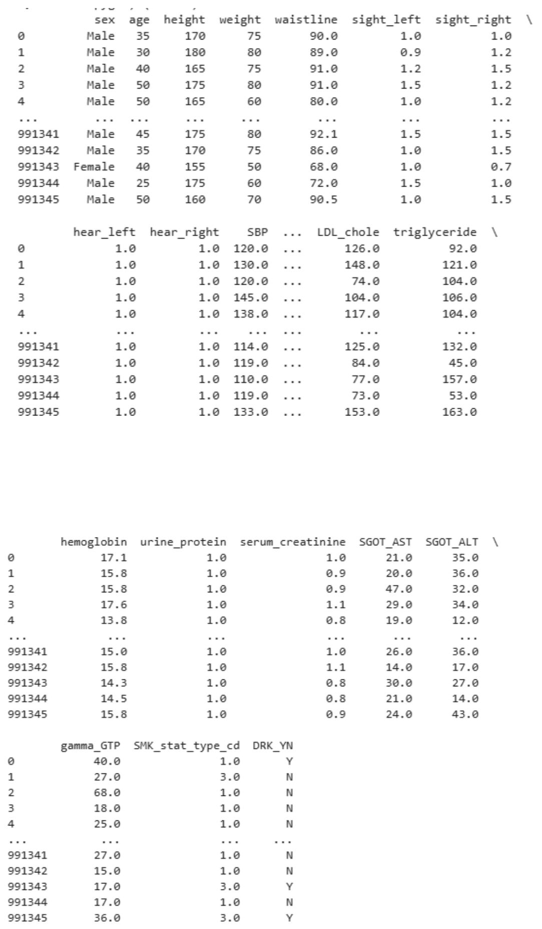
3. PROPOSED WORK AND RESULTS

3.1 Application Description

The proposed application predicts two key behaviors: drinking status (classified as “Drinker” or “Non-Drinker”) and smoking status (“Smoker” or “Non-Smoker”). It is designed for use by healthcare associates, enabling them to input a user's demographic and physiological data, such as age, cholesterol levels, liver enzyme metrics, and vision data, to obtain real-time predictions.

3.2 Datasets

The dataset used in this project, titled "Smoking and Drinking Dataset with Body Signal," was sourced from Kaggle [11]. It contains approximately 991,000 records and includes a comprehensive set of demographic features (age, sex, height, and weight) and physiological features (waistline, cholesterol levels, triglycerides, blood pressure, gamma-GTP, hemoglobin levels, vision metrics, and serum creatinine). The dataset was cleaned and preprocessed to ensure consistency and accuracy, with missing values handled through median imputation and outliers treated using interquartile range (IQR) methods. The data was split into 80% for training and 20% for testing to evaluate model performance effectively. This rich dataset provides a robust foundation for predicting smoking and drinking behaviors using machine learning models.



```
             sex  age  height  weight  waistline  sight_left  sight_right  \
0           Male   35     170      75       90.0         1.0          1.0
1           Male   30     180      80       89.0         0.9          1.2
2           Male   40     165      75       91.0         1.2          1.5
3           Male   50     175      80       91.0         1.5          1.2
4           Male   50     165      60       80.0         1.0          1.2
...          ...  ...     ...     ...        ...         ...          ...
991341      Male   45     175      80       92.1         1.5          1.5
991342      Male   35     170      75       86.0         1.0          1.5
991343    Female   40     155      50       68.0         1.0          0.7
991344      Male   25     175      60       72.0         1.5          1.0
991345      Male   50     160      70       90.5         1.0          1.5

        hear_left  hear_right    SBP  ...  LDL_chole  triglyceride  \
0             1.0         1.0  120.0  ...      126.0          92.0
1             1.0         1.0  130.0  ...      148.0         121.0
2             1.0         1.0  120.0  ...       74.0         104.0
3             1.0         1.0  145.0  ...      104.0         106.0
4             1.0         1.0  138.0  ...      117.0         104.0
...           ...         ...    ...  ...        ...           ...
991341        1.0         1.0  114.0  ...      125.0         132.0
991342        1.0         1.0  119.0  ...       84.0          45.0
991343        1.0         1.0  110.0  ...       77.0         157.0
991344        1.0         1.0  119.0  ...       73.0          53.0
991345        1.0         1.0  133.0  ...      153.0         163.0

        hemoglobin  urine_protein  serum_creatinine  SGOT_AST  SGOT_ALT  \
0             17.1            1.0               1.0      21.0      35.0
1             15.8            1.0               0.9      20.0      36.0
2             15.8            1.0               0.9      47.0      32.0
3             17.6            1.0               1.1      29.0      34.0
4             13.8            1.0               0.8      19.0      12.0
...            ...            ...               ...       ...       ...
991341        15.0            1.0               1.0      26.0      36.0
991342        15.8            1.0               1.1      14.0      17.0
991343        14.3            1.0               0.8      30.0      27.0
991344        14.5            1.0               0.8      21.0      14.0
991345        15.8            1.0               0.9      24.0      43.0

        gamma_GTP  SMK_stat_type_cd  DRK_YN
0            40.0               1.0       Y
1            27.0               3.0       N
2            68.0               1.0       N
3            18.0               1.0       N
4            25.0               1.0       N
...           ...               ...     ...
991341       27.0               1.0       N
991342       15.0               1.0       N
991343       17.0               3.0       Y
991344       17.0               1.0       N
991345       36.0               3.0       Y
```

Fig 1: Dataset Sample

3.3 System Architecture

The system architecture for this project integrates several key components to predict smoking and drinking behaviors efficiently and accurately. First, the **data preprocessing [Fig 2]** phase ensures data quality by cleaning and normalizing the dataset, scaling input variables for standardization, and encoding categorical features such as sex and smoking status. Next, **feature selection [Fig 6 and 7]** is performed using Decision Trees to rank variables by their importance. For drinking behavior predictions, key features include gamma-GTP, sex, age, hemoglobin, and HDL cholesterol, while smoking behavior predictions rely on gamma-GTP, waistline, serum creatinine, triglycerides, and age. The system employs multiple **machine learning models [3.4]**, including Logistic Regression, Decision Trees, XGBoost, and MARS, with XGBoost emerging as the best-performing model for drinking behavior (70% accuracy) and Logistic Regression excelling in smoking behavior predictions (68% accuracy). Finally, a graphical user interface (GUI) was developed using Python and HTML/CSS, enabling real-time data input and delivering clear, actionable predictions. This streamlined architecture ensures high accuracy and practical usability in healthcare applications.

3.3.1 Data Preprocessing

The data preprocessing phase was essential to ensure the dataset was clean, balanced, and ready for analysis. Missing values**[Fig 2]**, such as implausible waistline measurements of 999 cm, were replaced and imputed to maintain data integrity. Outliers**[Fig 4 and 5]** were treated using methods like the Interquartile Range (IQR), with extreme values in variables such as sight_left and sight_right capped at a maximum threshold of 4. Continuous variables, including blood pressure and cholesterol levels, were standardized and normalized to align scales for better model performance. Categorical variables, such as gender and smoking/drinking statuses, were converted into numerical formats using label encoding. Additionally, feature selection techniques identified key predictors, such as gamma-GTP, sex, and age for drinking, and healthcare access and medication use for smoking behavior. These steps enhanced the dataset's quality, ensuring robust and accurate model predictions.

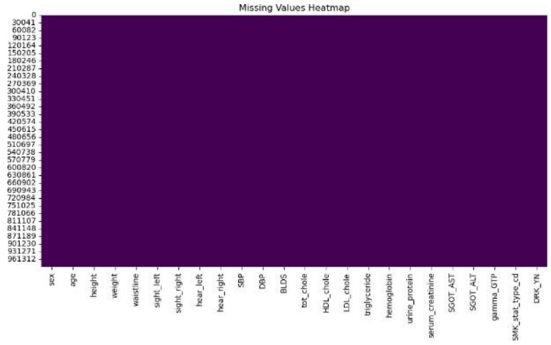


Fig2: Missing values Heatmap

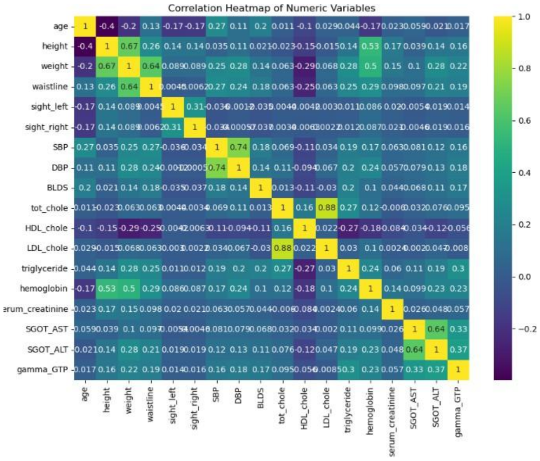
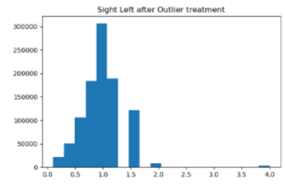


Fig 3:Correlation Heatmap of Numeric Values



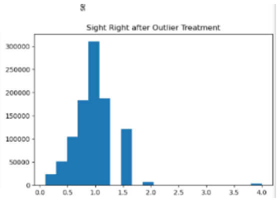


Fig4: Sight Left after Outlier treatment

Fig5: Sight Right after Outlier Treatment

3.3.2 Feature Selection

Feature selection is a vital step to identifying the most significant predictors in the dataset, improving model accuracy and interpretability. For drinking behavior **[Fig 6]**, key features such as gamma-GTP, sex, age, HDL cholesterol, and hemoglobin levels were identified as having the highest importance. Similarly, for smoking behavior **[Fig 7]**, significant predictors included healthcare access, medication use, waistline, gamma-GTP, and triglycerides, reflecting both physiological and socio-environmental factors. Decision Trees were employed to rank these features based on their contribution to the target variables, ensuring the removal of irrelevant or redundant data. This process enhanced the model's efficiency and focused on meaningful predictors, leading to more accurate and reliable results.

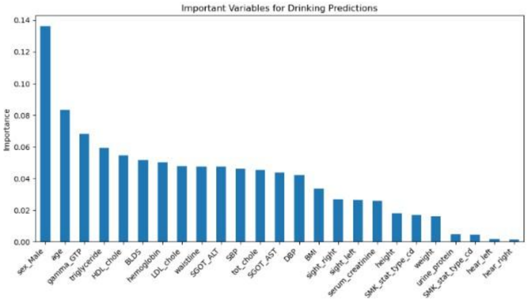


Fig6: Important Variables for Drinking Predictions

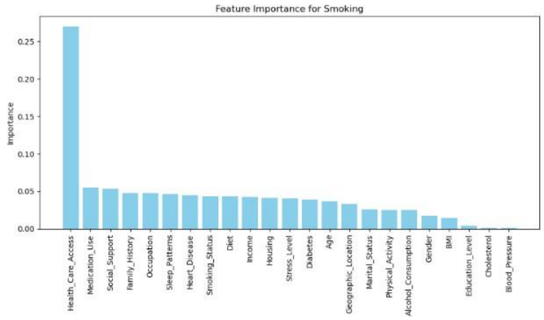


Fig 7: Feature Importance for Smoking

3.4 Algorithms

3.4.1.Logistic Regression Model

Logistic Regression**[6]** is a statistical method used to predict binary or multi-class outcomes by modeling the probability of a dependent variable belonging to a specific class. In this project, it is applied to classify drinking behaviors as “Drinker” or “Non-Drinker” and smoking statuses as “Smoker” or “Non Smoker” [Jawa, 2022]. It uses features such as gamma-GTP, hemoglobin levels, age, and waistline to generate probabilistic outputs that effectively distinguish between these behaviors.

$$p_i = f(y|x) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_i)\}} \qquad (1)$$

where p_i is the probability of success and, $P(y_i = 1) = p_i \text{and} P(y_i = 0) = q_i = 1 - p_i, 0 \leq p_i \leq 1.$

3.4.2 Decision Tree Model

Decision Trees[7] are supervised learning models that create a tree-like structure to classify data based on feature splits. For this project, Decision Trees are applied to rank feature importance and classify behaviors [Song, 2015]. Key features such as gamma-GTP, HDL cholesterol, waistline, and age are used to build the tree, which partitions the dataset into subsets based on the most informative splits. This model is particularly effective for identifying significant predictors for smoking and drinking behaviors.

3.4.3 MARS Model (Multivariate Adaptive Regression Splines)

MARS[8] is a non-parametric regression technique that builds models using piecewise linear regressions, capturing non-linear relationships and interactions between features[S Zurimi and Darwin 2020]. In this project, MARS is applied to model the complex relationships between physiological and demographic features, such as gamma-GTP, hemoglobin, and serum creatinine, and smoking and drinking behaviors.

3.4.4 XGBoost Model

XGBoost[9] is an advanced gradient boosting algorithm that builds decision trees sequentially, correcting errors from previous iterations [Nalluri, 2020]. In this project, XGBoost is utilized to classify both smoking and drinking behaviors by capturing complex interactions between features like gamma-GTP, HDL cholesterol, and triglycerides.

3.5 User Interface:

The application’s user interface allows health associates to enter a user’s demographic and physiological details, including Firstname, Lastname, email, DOB, age, sex, height, weight cholesterol levels, and other health indicators. Once the data is entered, it is submitted for processing by the predictive models. The system analyzes the inputs and generates predictions for drinking behavior (Drinker/Non-Drinker) and smoking behavior (Smoker/Non-Smoker). The results are displayed clearly in a user-friendly format, enabling quick and accurate interpretation. This streamlined process ensures efficient data handling and actionable insights for health-related decision-making.

3.6 Results and Analysis

The evaluation of model performance identified XGBoost[9] and Logistic Regression[6] as the two best-performing models for predicting smoking and drinking behaviors. XGBoost achieved the highest accuracy at 70%, along with the best F1 scores (Macro Avg: 0.61, Weighted Avg:

0.70), making it the most effective for both binary and multi-class classification tasks. It also demonstrated efficiency with a training time of 4.19 seconds. Logistic Regression followed closely, achieving an accuracy of 68% and a Weighted Avg F1-Score of 0.68, showing strong performance, particularly in predicting smoking behavior. While Decision Tree[7] and MARS[8] models had accuracies of 62% and 61%, respectively, they fell short in precision and speed compared to the top two models.

Justifications for Model Choices:

XGBoost[9] was selected due to its superior accuracy (70%) and F1 scores (Macro Avg: 0.61, Weighted Avg: 0.70), making it the most effective model for both binary and multi-class classification tasks. Its efficiency, with a training time of 4.19 seconds, further supports its suitability. Logistic Regression[6] was also chosen for its strong performance in smoking behavior predictions, achieving 68% accuracy and a Weighted Avg F1-Score of 0.68.

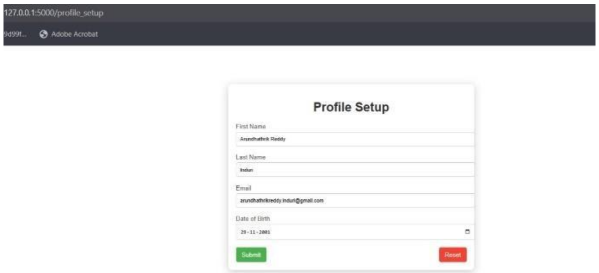
3.7 User Manual:

The GUI is designed for healthcare associates to predict a user's smoking and drinking behaviors efficiently. Users begin by launching the application and entering basic demographic details such as Firstname, Lastname, Email, Dob, age, sex, height, and weight, followed by physiological data including waistline, cholesterol levels, blood pressure, gamma-GTP, hemoglobin, and other health indicators. After ensuring all fields are filled correctly, the data is submitted for processing through the predictive models. The system then analyzes the inputs and provides clear predictions on the user's drinking status (Drinker or Non-Drinker) and smoking status (Smoker or Non-Smoker). These results are displayed in an intuitive format, ensuring accessibility and ease of interpretation for healthcare professionals.

3.8 Experiment:

The experiment involves a Health Assessment Application designed to evaluate user-provided health data and deliver insights. It starts with a Profile Setup, where users enter personal details like name, email, and date of birth. Next, the application collects Medical Details, including metrics like waistline, blood pressure, cholesterol levels, blood sugar, and sensory abilities (sight and hearing). These inputs are processed to provide health insights. In a Sample Experiment, We entered specific values, and the system classified me as a "Smoker" and "Alcoholic" based on the submitted data. This demonstrates the application's ability to analyze medical metrics and deliver personalized results effectively. The experiment highlights the application's potential to leverage user data for actionable health insights, showcasing its role in personalized health analytics

3.9 Screenshots of prediction results:



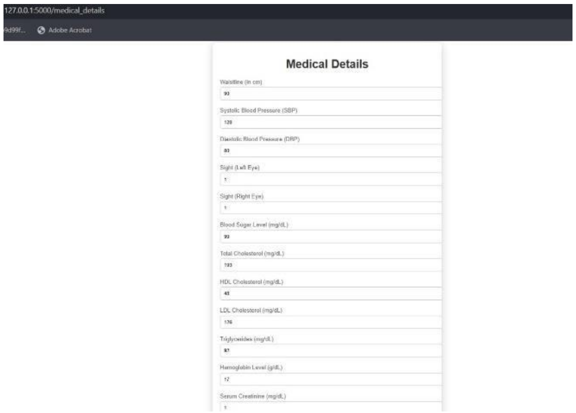
Profile Setup

First Name: Arundhathik Reddy

Last Name: Indurі

Email: arundhathikreddy.induri@gmail.com

Date of Birth: 29-11-2001



127.0.0.1:5000/medical_details

Medical Details

Waistline (in cm)
93

Systolic Blood Pressure (SBP)
120

Diastolic Blood Pressure (DBP)
80

Sight (Left Eye)
1

Sight (Right Eye)
1

Blood Sugar Level (mg/dL)
93

Total Cholesterol (mg/dL)
193

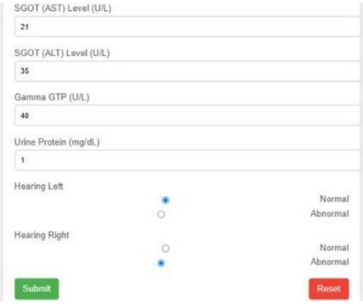
HDL Cholesterol (mg/dL)
45

LDL Cholesterol (mg/dL)
126

Triglycerides (mg/dL)
97

Hemoglobin Level (g/dL)
17

Serum Creatinine (mg/dL)
1



SGOT (AST) Level (U/L)
21

SGOT (ALT) Level (U/L)
35

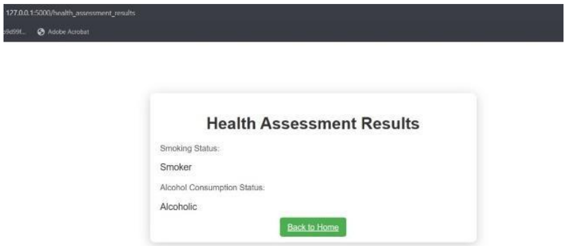
Gamma GTP (U/L)
40

Urine Protein (mg/dL)
1

Hearing Left
● Normal
○ Abnormal

Hearing Right
○ Normal
● Abnormal

Result:



127.0.0.1:5000/health_assessment_results

Health Assessment Results

Smoking Status:
Smoker

Alcohol Consumption Status:
Alcoholic

4. CONCLUSIONS AND FUTURE WORK

The project on predicting smoking and drinking behaviors using physiological and demographic data has successfully laid the groundwork for utilizing machine learning to address real-world health and insurance challenges. By applying advanced classification models like Logistic Regression[6], Decision Trees[7], MARS[8], and XGBoost[9], the project achieved meaningful insights into behavioral prediction, with XGBoost emerging as a highly effective model. The preprocessing and feature selection processes ensured robust and clean data for analysis, enhancing the reliability of results. Future work could expand the scope by incorporating longitudinal datasets to study behavioral trends over time, integrating additional physiological markers for more nuanced predictions, and refining the user interface to enhance accessibility and interactivity. Furthermore, extending the application to include real-time data collection through wearable devices could transform it into a dynamic health-monitoring tool. Continuous evaluation and model optimization will remain crucial for maintaining relevance and accuracy, ensuring the system's practical adoption in healthcare and insurance industries.

REFERENCES

[1] Sunday Azagba, Neill Bruce Baskerville, and Leia Minaker. “A comparison of adolescentsmoking initiation measures on predicting future smoking behavior.” Preventive Medicine Reports, vol 2, 2015, pages 174-177.

[2] Xiang Zhao, Franziska F.Dichtl and Heather M. Foran. “Predicting Smoking Behavior:intention and future self-continuity among Austrians.” Psychology, Health and Medicine, vol27, 2022-Issue 5.

[3] Richard Cooke, Falko Sniehotta and Benjamin Schuz. “Predicting Binge-

Drinking Behaviour using an extended TPB: Examining the Impact of Anticipated Regret and Descriptive Norms." Alcohol and Alcoholism, vol 42, Issue 2, March 2007, Pages 84-91.

[4] Wenbin Liang, Huijun Chih and Tanya Chikritzhs. "Predicting Alcohol Consumption Patterns for Individuals with a User- Friendly Parsimonious Statistical Model." International Journal of Environmental Research and Public Health. 2023 Feb; 20(3): 2581.

[5] Charles Frank, Asmail Habach and Raed Seetan. "Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis." Research Gate, March 2018. Advances in Science Technology and Engineering Systems Journal.

[6] Taghreed M. Jawa. "Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia."Alexandra Engineering Journal. Volume 61, Issue 10, October 2022, Pages 7995-8005.

[7] Yan-yan SONG and YingLU. " Decision tree methods:applications for classification and prediction." National Library of Medicine. Shanghai Arch Psychiatry. 2015 Apr 25;27920:130-135.

[8] S Zurimi and Darwin. "Analysis of Multivariate Adaptive Regression Spline (MARS) Model in Classifying factors affecting on Student the Study Period at FKIP Darussalam University of Ambon."Journal of Physics: Conference Series, Volume 1463, The 5th International Conference on Basic Sciences 5-6 September 2019.

[9] Mounika Nalluri and Nageshwar rao Eluri. " A Scalable Tree Boosting system: XGBoost."Research Gate , January 2020.International Journal of Research Studies in Science,Engineering and Technology Volume 7, Issue 12, 2020, PP 36-51.

[10] Volkan Y. Senyureka, Masudul H. Imtiaz a, Prajakta Belsarea, Stephen Tiffany b and Edward Sazonova. "Smoking detection based on regularity analysis of hand to mouth gestures."Biomedical Signal Processing and Control, Volume 51, May 2019, Pages 106-112.

[11] Dataset Link:

[https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset/data