ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

LEVERAGING MACHINE LEARNING FOR HEALTH IMPACT ANALYSIS OF SOFT DRINK CONSUMPTION USING ENSEMBLE METHODS

[1]Michael Chukwuemeka Okafor and [2]Adedayo Temidayo Adewale

[1]Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria.

[2]Faculty of Information Technology, Akwa Ibom State University, Mkpat Enin, Nigeria.

DOI: 10.5281/zenodo.14892607

Abstra

Soft drinks, often high in added sugars, have raised significant public health concerns due to their association with adverse health effects such as obesity, type 2 diabetes, and cardiovascular diseases. The aim of this study is to investigate the health impact of soft drink consumption, particularly focusing on the implications of excessive sugar intake. To address this concern, we employed four Ensemble Learning approaches to assess the health risks associated with soft drink consumption, specific Ensemble Learning approaches such as LightGBM, CatBoost, XGBoost, and Random Forest we adopted. Our findings indicate that older individuals may not require as much soft drink consumption, and the general population should limit their daily intake to a single bottle, approximately 35g per day. The study reveals that using Ensemble Learning techniques, including LightGBM, CatBoost, XGBoost, and Random Forest, yielded promising results, with CatBoost emerging as the top-performing model with a model accuracy score of 97%, surpassing the performance of Random Forest and XGBoost algorithms. The research reveals the importance of limiting the consumption of sugary beverages and opting for healthier alternatives to mitigate the risk of adverse health outcomes associated with excessive sugar consumption. Overall, our findings contribute valuable insights into understanding the health implications of soft drink consumption and highlight the efficacy of Ensemble Learning approaches in assessing and addressing public health concerns. This study provides actionable recommendations for individuals and health organizations to promote healthier dietary habits and mitigate the risk of sugar-related health complications, emphasizing the importance of evidence-based interventions in safeguarding public health.

Keywords: Sugar, Soft drinks, Ensemble Learning, Health, Diabetes.

1. INTRODUCTION

The consumption of soft drinks, particularly sugary beverages, has become a widespread habit in many parts of the world (Esrafil et al., 2022). While these drinks are often enjoyed for their taste and convenience, there is growing concern about their potential health impacts(Doris et al., 2019). Soft drinks are typically high in added sugars, which have been associated with various adverse health effects, including obesity, type 2 diabetes, cardiovascular diseases, and dental problems. As a result, the excessive consumption of soft drinks has raised significant public health concerns, prompting researchers to explore innovative approaches to assess their health impact (Esrafil et al., 2022). Traditional epidemiological studies have provided valuable insights into the associations between soft drink consumption and adverse health outcomes. However, these studies

often rely on observational data, which may be subject to various biases and limitations (Ekong et al., 2023) Moreover, the complex interplay of multiple factors contributing to health outcomes makes it challenging to establish causal relationships definitively (Azuma et al., 2020). Machine learning, a subset of artificial intelligence, offers a promising avenue for addressing these challenges. By leveraging advanced computational techniques, machine learning can analyze vast datasets, identify patterns, and generate predictive models. These models can help researchers and healthcare professionals better understand the intricate relationships between soft drink consumption and health outcomes (Doris et al., 2019). Multi-Ensemble algorithms, which draw inspiration from natural processes and phenomena, have gained recognition for their ability to solve complex optimization problems (Umoren & Inyang, 2021). These algorithms mimic the behavior of biological systems, such as genetic evolution, swarm intelligence, and neural networks, to find optimal solutions in diverse domains (Joseph et al., 2022). In the context of soft drink consumption and its health impact, Multi-Ensemble algorithms can be employed to model and simulate the intricate biological processes within the human body (Azuma et al., 2020). By simulating how the consumption of sugary beverages affects various physiological parameters, these algorithms can help predict potential health risks associated with soft drink intake. This holistic approach enables researchers to consider multiple variables simultaneously and capture the dynamic interactions between them (Ekong et al., 2022). The research on a machine learning system for health impact assessment of soft drink consumption using Multi Ensemble learning approach represents an innovative and multidisciplinary approach to address the complex issue of soft drink-related health risks (Azuma et al., 2020). By combining the analytical power of machine learning with Multi Ensemble optimization techniques, this research endeavors to provide a more comprehensive understanding of the health consequences of soft drink consumption and, ultimately, contribute to informed dietary choices and public health interventions . XGBoost, LightGBM, Random Forest, and CatBoost are all versatile machine learning algorithms(Inyang & Umoren, 2023) that are well-suited for conducting a health impact assessment of soft drinks on human health. Their collective strengths make them effective choices for this task (Inyang & Umoren, 2023). These algorithms are proficient at handling complex and high-dimensional data, which is essential when assessing the nutritional and ingredient profiles of various soft drinks. For instance, they can analyze the relationships between different ingredients and health outcomes, such as the correlation between high sugar content and obesity, or the impact of artificial additives on various health conditions. Furthermore, they can provide feature importance scores, helping researchers identify the most influential factors contributing to health impacts. This can guide health professionals and policymakers in making informed decisions about soft drink consumption and health policies. XGBoost, LightGBM, Random Forest (Edet et al., 2024), and CatBoost are also adept at classification tasks (Edet & Ansa, 2023), making them suitable for categorizing soft drinks based on their ingredients and nutritional content. By classifying soft drinks into groups such as "healthy" or "unhealthy," these algorithms can assist in providing clear guidelines for

consumers and facilitate public health awareness campaigns. Today, our world boasts a multitude of soft drink manufacturers, each producing various brands of beverages. Within these drinks, certain ingredients are present in substantial quantities, particularly sugar and carbohydrates, which can metabolize into sugar within the human body. The persistent consumption of such products can potentially yield significant adverse effects on the health of those who consume them. In this research endeavor, we aim to conduct a comprehensive analysis of the repercussions of these products on human health. The significance of investigating the health impact of soft drink consumption cannot be overstated. Soft drinks have become ubiquitous in our modern society, with widespread availability and consumption across age groups and demographics. However, the contents of these beverages, often high in sugar and carbohydrates, have been linked to various health issues, including obesity, type 2 diabetes, and cardiovascular diseases. Understanding the precise health implications of soft drink consumption is crucial as it directly affects public health policies, dietary recommendations, and personal choices. Moreover, this research delves into a unique approach by employing Multi-Ensemble Learning algorithms, which draw inspiration from natural processes and systems(Uwah & Edet, 2024), to assess the health consequences of soft drink consumption. This interdisciplinary approach holds the promise of revealing nuanced insights that traditional analyses might miss.

2. LITERATURE REVIEW

Esrafil et al., (2022) Proposed a work on Identification and quantification of sodium benzoate in soft dripnks available in Tangail region by high-performance liquid chromatography. In the study, an experimental work was performed by high-performance Liquid Chromatography (HPLC) in order to determine the quantity of sodium benzoate in various brand of soft drinks available in the markets , stores and shops in the Tangail region of Bangladesh. The weakness in this work is that the author did not state the impact of consuming this soft drinks in human system, hence, there is a need for improvement. Akolawole et al., 2022 proposed a work on Effect of storage on the levels of sodium benzoate in soft drinks sold in some Nigerian market with exposure and health risk assessment. In all, fifty (50) soft drinks samples, acquired from Enugu, Aba, Asaba, Onitsha and Owerri markets in Nigeria, were subjected to four different storage conditions namely: room temperature (RT), refrigerated (RF), 40 °C and 60 °C for 15 days after which they were analyzed for sodium benzoate concentration using HPLC – UV/Vis detector. The results showed on the average that at RT and RF, soft drinks from Aba had the highest concentration of sodium benzoate (98.7 mg/L and 112.9 mg/L) respectively while samples from Asaba had least concentration of 39.9 mg/L and 38.1mg/L. At increased temperature of 40 °C, the concentration of sodium benzoate increased generally across the sample, while at 60 °C, the levels in all the samples analyzed were either reduced to less than 50% or below detection level, which suggest that degradation of sodium benzoate at this elevated temperature could result in benzene formation, which is a known carcinogen. Carcinogenic and non-carcinogenic risk assessment showed that children are at risk compared to adults due to higher sodium benzoate daily intake leading to high rate of hyperactivity in

correlation to malaise.The weakness is that the work does not indicate the effect of additives such as preservatives on human health.Tahmassebi and BaniHani, (2020), conducted a critical review of the literature using electronic databases, Medline, Embase, Cochrane library, and cross-referencing using published references. They found that soft drink consumption can contribute to detrimental oral and general health. The consumption of soft drinks was found to have increased dramatically over the past several decades, particularly among children and adolescents. It was only the content of soft drinks that was analyzed, no preservatives was mentioned. The old or elderly people were not considered in the solution designed.The critical review by BaniHani et al. (2019) found that soft drink consumption can contribute to detrimental oral and general health. The consumption of soft drinks was found to have increased dramatically over the past several decades. Some commercial soft drinks are high in sugar content and acidity, and therefore, their consumption can contribute to detrimental oral and general health. The work did not consider the effect of preservatives on human health, the emphasis was on the content of the drinks.The study by Philipp et al. (2016) found that soft drink consumption is associated with mental health problems in children and adolescents. Soft drink consumption is considered a major risk factor for the development of widespread noncommunicable diseases such as obesity and type 2 diabetes. According to Harvard T.H. Chan School of Public Health, 2023, sugar-sweetened soft drinks contribute to the development of diabetes and cardiovascular disease. Each additional serving per day of sugary drink was linked with a 10% increased higher risk of cardiovascular disease-related death. The more ounces of sugary beverages a person has each day, the more calories he or she takes in later in the day.The systematic review and meta-analysis by Vartanian et al.(2021) found clear associations of soft drink intake with increased energy intake and body weight, as well as lower intakes of milk and calcium. Soft drink consumption was also associated with an increased risk of chronic diseases.The study by De Koning et al.(2023) found that sweetened beverage consumption is associated with an increased risk of coronary heart disease in men. The review by Brown and Rother, (2023), discusses the role of artificial sweeteners in soft drinks and their potential impact on health. The authors suggest that more research is needed to fully understand the health effects of artificial sweeteners. The review by Popkin et al.(2023) discusses the impact of soft drink consumption on global health. The authors suggest that reducing soft drink consumption may be an effective strategy for preventing a range of non-communicable diseases, including obesity, type 2 diabetes, and cardiovascular disease. Hu et al. (2023) proposed a work on Consumption of Soft Drinks and Overweight and Obesity Among Adolescents in 107 Countries and Regions. The aim of the authors was to investigate the association of soft drink consumption with overweight and obesity in adolescents enrolled in school (hereafter, school-going adolescents) using country-level and individual level data.However, the work did not consider what would happen if the preservatives used in preserving the drinks are considered in the design of the problem. In this work, we wish to incooperate the preservatives. The study by Rubaiyai et al. (2022) developed an image-based soft drink type classification and dietary assessment system using deep convolutional neural

network with transfer learning. The system was able to recognize and classify different types of drinks and estimate their calorie content. The authors suggest that this system can be used to monitor and assess the dietary intake of individuals and help fight overweight and obesity. This system was a tool designed to detect types of soft drinks and predict their calorie content. The weakness in this system is that it does not check for the effect of the preservatives used on human body. The system also failed to predict the kind of diseases one could get by consuming it.

3. METHOD

In this work, the Ensemble Learning is used. Ensemble Learning approaches such as the Random Forest, XGBoost, CatBoost and LightGBM are adopted for analysis and assessment of soft drink data to determine its impact on human health. The algorithmic flow and the mathematical representation of the model is presented.

A. DATA COLLECTION

This research aims to determine the health risks in consuming more than the recommended number of 35cl Cola products in a day. Determining how many bottles of Cola one can consume in a day without encountering sugar-related issues depends on various factors, including individual health, activity level, overall diet, and any existing health conditions such as diabetes or obesity. However, health organizations such as the World Health Organization (WHO) recommend limiting added sugar intake to no more than 10% of total daily calories. For an average adult with a daily caloric intake of around 2000-2500 calories, this would mean limiting added sugar intake to approximately 50 grams per day. Given that a 330ml bottle of Coca-Cola Classic contains around 35 grams of sugar, consuming one bottle would already account for a significant portion of the recommended daily sugar intake. Consuming multiple bottles in a day could easily exceed the recommended limit. Therefore, it's generally advisable to limit consumption of sugary beverages like Cola and to opt for healthier alternatives such as water, unsweetened tea, or sparkling water flavored with natural ingredients. The sample used in this work focused on specific features such as :

`Product`: Indicates the type of soft drink (in this case, cola).

`Bottles_Daily`: Represents the number of bottles of cola consumed by an individual in a day.

`Sugar_Per_Bottle(g)`: Amount of sugar in grams per bottle of cola (assuming a 35cl bottle of Cola product).

`Total_Sugar(g)`: Total amount of sugar consumed in grams based on the number of bottles consumed.

`Recommended_Bottle_Daily`: The recommended maximum number of bottles of cola to consume daily, set to 1 for this example.

`Risk`: Indicates whether the total sugar consumption exceeds the recommended daily limit (Yes/No). In this case, consuming any amount of cola beyond one bottle daily poses a risk according to the dataset's criteria.

B. DATA PREPROCESSING

Data preprocessing is the initial stage in the data analysis workflow, encompassing various operations to refine and format raw data for subsequent analysis or modeling. In process we focused on tasks like handling missing

values, identifying and rectifying errors, transforming data into a suitable representation, and selecting relevant features. By addressing inconsistencies and noise in the data, preprocessing enhances the quality and reliability of this analyses, facilitating more accurate and meaningful insights.

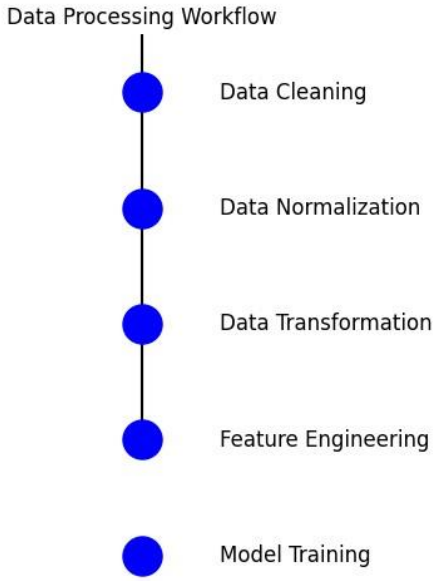


Fig.1.0 Data Preprocessing Workflow

Through techniques such as normalization, encoding, data preprocessing streamlined the dataset, making it more manageable and conducive to efficient analysis. It has been observed that effective data preprocessing lays the groundwork for robust and insightful data-driven decisions.

Table 1.0 shows the features and their respective data types in the dataset:

Table 1: Dataset Attributes and DataType

Attribute	**Data Type**
Product	Categorical
Bottles_Daily	Integer
Sugar_Per_Bottle(g)	Integer/Float
Recommended_Bottle_Daily	Integer
Age	Integer
BMI	Float
Total_Sugar(g)	Integer/Float
sugar_related_issue	Integer

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

Risk	Categorical

This table, Table 1.0, provides an overview of the features and their corresponding data types. Categorical data types represent discrete categories, while integer and float data types represent numerical values.

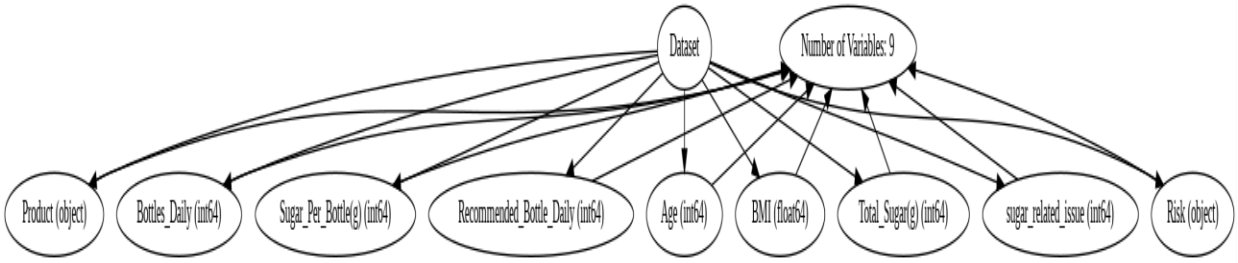


Fig. 2.0 Dataset Structure

Figure 2.0 shows the structure of the dataset used in this research. It shows the dataset and the variables that form the dataset. The datatype for each of the variables are shown to ensure error free computation.

	Product	Bottles_Daily	Sugar_Per_Bottle(g)	Recommended_Bottle_Daily	Age	BMI	Total_Sugar(g)	sugar_related_issue	Risk
0	Cola	2	35	1	50	24.326752	70	0	Yes
1	Cola	1	35	1	35	25.744865	35	1	No
2	Cola	3	35	1	24	26.868286	105	1	Yes
3	Cola	2	35	1	32	19.298289	70	0	Yes
4	Cola	2	35	1	19	30.349312	70	1	Yes

Fig. 3.0 Snapshot of the Processed Dataset

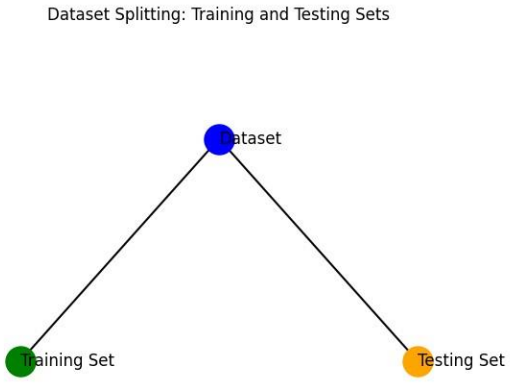


Fig. 4.0 Data Splitting

In Figure 4.0, the dataset is represented at the center, with the training set on the left and the testing set on the right. Arrows indicate the flow of data from the dataset to each set. In data splitting, the dataset is divided into

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

two subsets: the training set and the testing set. The training set is used to train the machine learning model, where it learns patterns and relationships between the input features and the target variable through various algorithms and optimization techniques. This process involves adjusting the model's parameters to minimize the difference between the predicted outcomes and the actual values in the training data. Once the model is trained, the testing set is utilized to evaluate its performance and generalization ability. The testing set serves as an independent dataset that the model has not seen during training, allowing for an unbiased assessment of its predictive accuracy. By splitting the data into training and testing sets, researchers can effectively train and validate machine learning models, ensuring that they can accurately predict outcomes on unseen data, thereby enhancing their reliability and applicability in real-world scenarios.

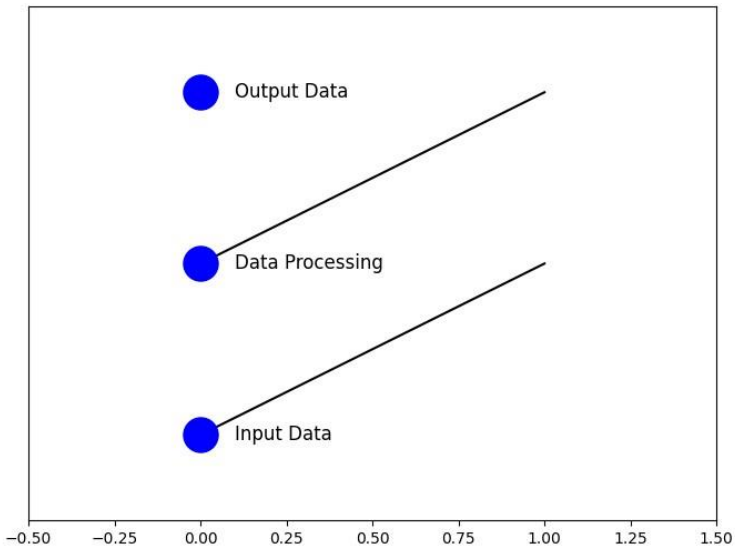


Fig. 5.0 Computational Segments

In any data processing workflow, there are three fundamental segments or processes: input, processing, and output. Starting from the bottom of the pipeline, the input segment is where raw data enters the system. This raw data can originate from various sources such as databases, sensors, user inputs, or external feeds, in this case, our dataset. Once the data is ingested, it moves on to the processing segment where it undergoes a series of transformations, analyses, and computations. This phase involves cleaning and formatting the data, extracting relevant information, and applying algorithms or models to derive insights or perform specific tasks. Finally, in the output segment, the processed data is presented in a usable format to end-users, systems, or other downstream processes. This could involve generating reports, visualizations, alerts, or providing data for further analysis or decision-making. The structured flow from input to processing to output ensures that data

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

is efficiently handled and transformed into valuable insights or actions, thereby facilitating informed decision-making and driving business outcomes.

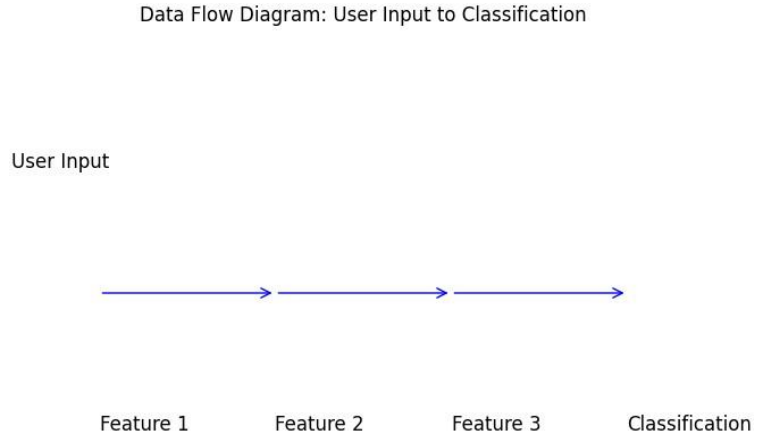


Fig.6.0 Input Flow for Classification

In the data processing workflow, data flows seamlessly from the input segment through to the final output following processing. This structured flow is depicted in the diagram, illustrating how each feature's data values are entered and processed to enable the classification of health risk. Starting at the input segment, raw data is received from various sources and ingested into the system. Subsequently, the data undergoes processing where it is cleaned, transformed, and analyzed to extract meaningful insights. During this phase, each feature's data values are examined and utilized in conjunction with algorithms or models to classify health risk accurately. Finally, the processed data flows to the output segment where the results of the classification process are presented to end-users or systems for further analysis or action. This systematic data flow ensures that information is effectively processed and utilized to drive informed decision-making regarding health risk assessment.

C. MODEL AND PROBLEM FORMULATION

The selected Ensemble Learning approaches, XGBoost, LightGBM, Random Forest, and CatBoost are adept at classification tasks, making them suitable for categorizing soft drinks based on their ingredients and nutritional content. By classifying soft drinks into groups such as "healthy" or "unhealthy," these algorithms can assist in providing clear guidelines for consumers and facilitate public health awareness campaigns. Here, the general problem solving flow for each Ensemble Learning techniques used is presented:

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

I. Algorithmic flow for Random Forest

A random forest is an ensemble machine learning technique used for both classification and regression tasks. It is a powerful and versatile algorithm that combines multiple decision trees to make more accurate and robust predictions.

The mathematical model of a random forest can be represented as follows:

Let N be the number of decision trees in the forest.

Each decision tree i can be represented as f_i(x), where x is the input data.

The ensemble prediction is obtained by majority voting:

[{y}(x) = Sum{i=1}^{N}{I}(f_i(x) = c)] Where:

{y}(x) represents the ensemble prediction for input x. f_i(x) represents the prediction of the i-th decision tree for input x. N is the total number of decision trees in the random forest.

c represents the class labels in the case of a classification problem.

{I} is the indicator function that evaluates to 1 if the condition in the parentheses is true and 0 otherwise.

II. Algorithmic flow for XGBoost

XGBoost (Extreme Gradient Boosting) is another popular ensemble machine learning algorithm, particularly effective for both regression and classification tasks. Below is the mathematical representation of the XGBoost model:

Objective Function for XGBoost

XGBoost aims to optimize an objective function that measures the model's performance. For regression tasks, the objective function is typically the mean squared error (MSE):

[{MSE} = {1}{N}sum_{i=1}^{N}(y_i {y}_i)^2] where:

(N) is the number of data points.

(y_i) is the true target value for data point (i).

({y}) is the predicted value for data point (i).

XGBoost Model

The XGBoost model can be represented as the sum of multiple decision trees:

[{y}(x) = \sum_{i=1}^{N_{trees}} f_i(x)] where:

(N_{trees\) is the number of trees in the XGBoost model.

(f_i(x)) represents the prediction of the (i)-th decision tree for input (x).

III. Algorithmic flow for CatBoost

CatBoost is another popular gradient boosting algorithm that is particularly effective for various machine learning tasks. Here's the mathematical representation of the CatBoost model

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

MSE = (1/N) * Σ(yi - ŷi)^2
Cross-Entropy Loss = -Σ(yi * log(p(yi)) + (1 - yi) * log(1 - p(yi))) Where:
N is the number of data points.
yi is the true target value for data point i. ŷi is the predicted
value for data point i.

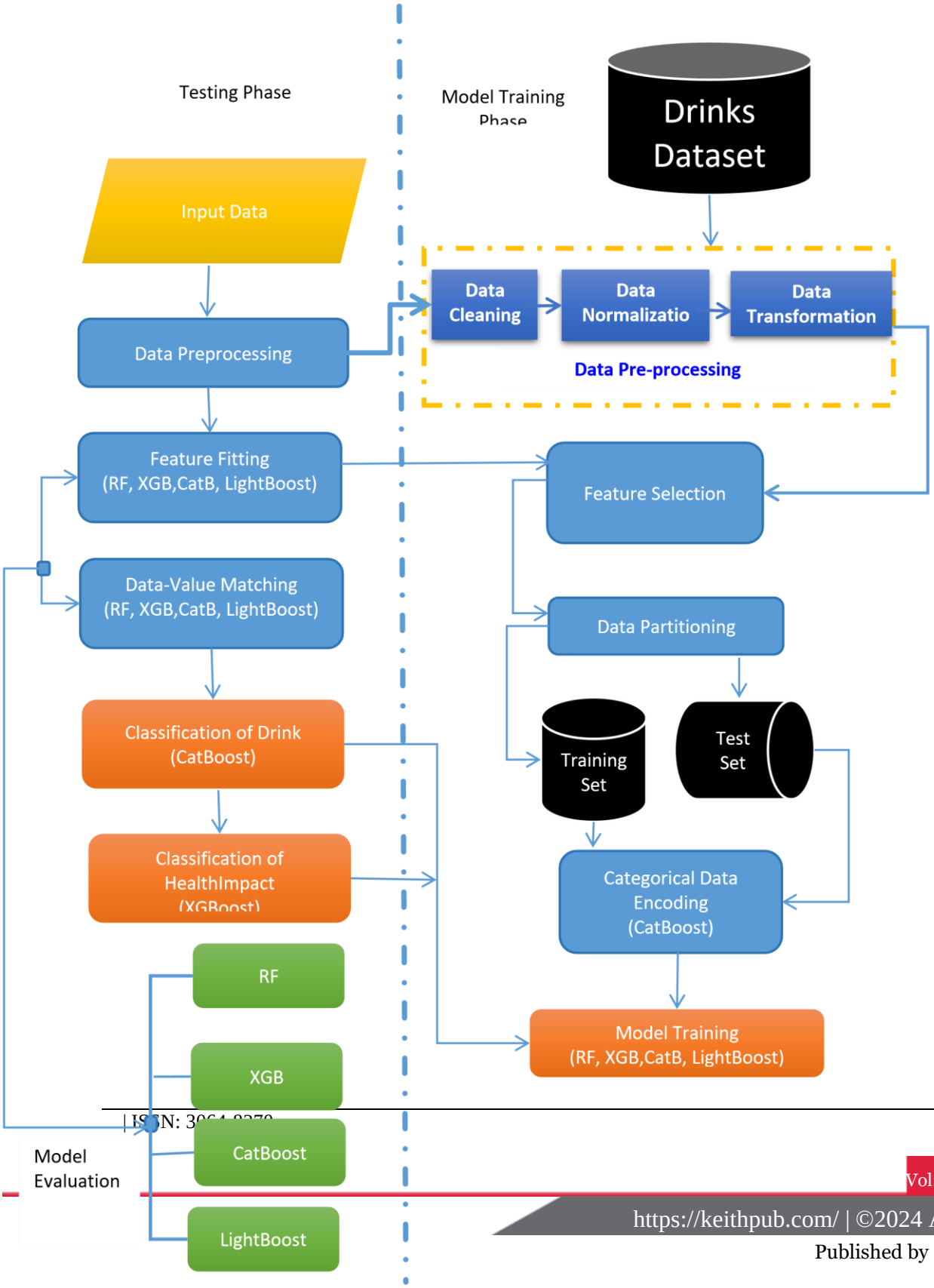
IV. Algorithmic flow for LightGBMoost

LightGBM is a gradient boosting framework developed by Microsoft that is known for its speed and efficiency. Here, we will provide a simplified mathematical representation of the LightGBM algorithm:
The LightGBM model can be represented as the sum of multiple decision trees: ŷ(x) = Σ f_i(x)
p(class_1 | x) = Σ p_i(class_1 | x)

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article



ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

Fig.7.0 Conceptual Framework

4. RESULTS AND DISCUSSION

In this section, we present, discuss, and document the results, research findings, and overall outcomes of the study. Through a comprehensive analysis of the data and application of various methodologies, including machine learning techniques and statistical analyses, we elucidate key insights and draw conclusions regarding the research objectives. The findings are contextualized within the broader scope of the study's aims and objectives, providing valuable insights into the underlying patterns, trends, and implications of the research, ensuring a thorough and rigorous examination of the research findings and their implications for theory, practice, and policy.

Table 2.0 Random Forest Model Report

Random Forest				
	precision	recall	f1-score	support
No	0.89	0.90	0.90	659
Yes	0.96	0.96	0.96	1570
accuracy			0.94	2229
macro avg	0.93	0.93	0.93	2229
weighted avg	0.94	0.94	0.94	2229

Table 2.0 presents the model summary of the Random Forest algorithm, showcasing an impressive accuracy score of 94%. This high accuracy indicates a robust level of performance in predicting outcomes based on the input data. The Random Forest algorithm demonstrates its efficacy in capturing complex relationships within the dataset, thereby yielding reliable predictions. The table serves as a concise yet informative overview of the model's performance, providing valuable insights into its predictive capabilities and reaffirming its suitability for the task at hand.

Table 3.0 XGBoost Model Report

XGBoost				
	precision	recall	f1-score	support
0	0.9	0.90	0.94	659
1	0.99	0.96	0.97	1570
accuracy			0.96	2229
macro avg	0.95	0.97	0.96	2229

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

weighted avg	0.97	0.96	0.96	2229

Table 3.0 offers a comprehensive report on the classification of soft drinks based on their sugar content and associated impact on human health, specifically utilizing the XGBoost algorithm. With an impressive accuracy score of 96%, XGBoost outperforms the Random Forest algorithm, showcasing its superior predictive capabilities in this context. The report encapsulates key metrics and insights derived from the classification process, providing valuable information on the model's performance and its implications for assessing the health implications of soft drink consumption. This detailed overview underscores the efficacy of XGBoost in accurately classifying soft drinks and highlights its potential utility in informing public health initiatives and interventions aimed at mitigating the adverse effects of high sugar consumption.

Table 4.0 CatBoost Model Report

CatBoost				
	precision	recall	f1-score	support
0	0.9	0.99	0.95	659
1	1	0.96	0.98	1570
accuracy			0.97	2229
macro avg	0.95	0.97	0.96	2229
weighted avg	0.97	0.97	0.97	2229

Table 4.0 showcases the classification reports generated by the CatBoost algorithm for analyzing soft drink sugar content data. With an impressive model accuracy score of 97%, CatBoost demonstrates superior strength, surpassing both the Random Forest and XGBoost algorithms in this analysis. The classification reports offer detailed insights into the model's performance metrics, highlighting its robust predictive capabilities and effectiveness in accurately classifying soft drinks based on their sugar content. This exceptional accuracy underscores CatBoost's efficacy as a powerful machine learning algorithm for addressing complex classification tasks and its potential to yield valuable insights into the health implications of soft drink consumption.

Table 5.0 LightGBM Model Report

LightGMB				
	precision	recall	f1-score	support
0	0.9	0.99	0.95	659

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

1	1	0.96	0.98	1570
accuracy			0.97	2229
macro avg	0.95	0.97	0.96	2229
weighted avg	0.97	0.97	0.97	2229

Table 5.0 presents the LightGBM model summary on the classification of the data used in this study. It has equal model accuracy score of 97% as the CatBoost Algorithm.

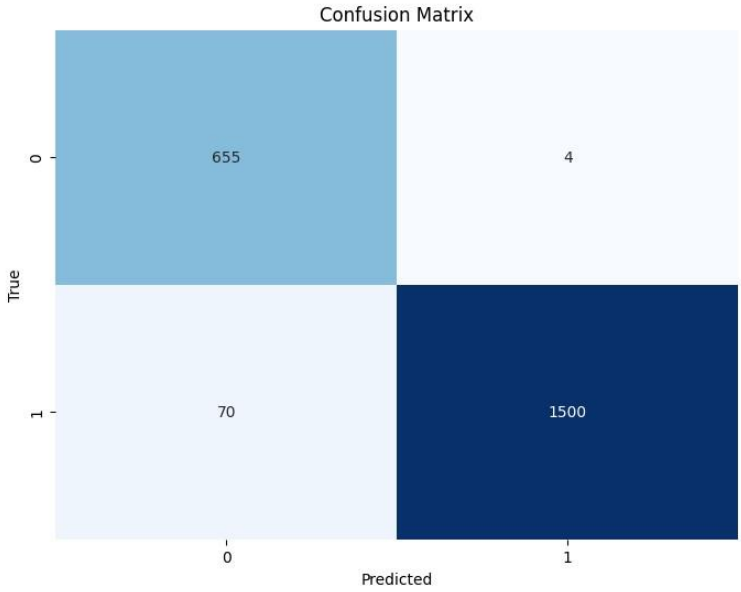


Fig.8(a) Confusion Matrix

The confusion matrix serves as a crucial tool for evaluating the performance of classification models in predicting the risk of sugar-related health complications. It allows us to analyze the accuracy of the model's predictions by comparing them to the actual outcomes observed in the data. By breaking down the classification results into true positive, true negative, false positive, and false negative predictions, the confusion matrix provides insights into the model's ability to correctly identify individuals at risk of sugar-related health issues and those not at risk. Through the confusion matrix, we assessed the model's sensitivity, specificity, precision, and overall accuracy in classifying individuals based on their soft drink consumption habits and associated health outcomes. This evaluation enables us to quantify the model's performance in identifying individuals at risk of sugar-related health complications, which is essential for informing targeted interventions and public health policies aimed at reducing the adverse effects of excessive soft drink consumption.

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

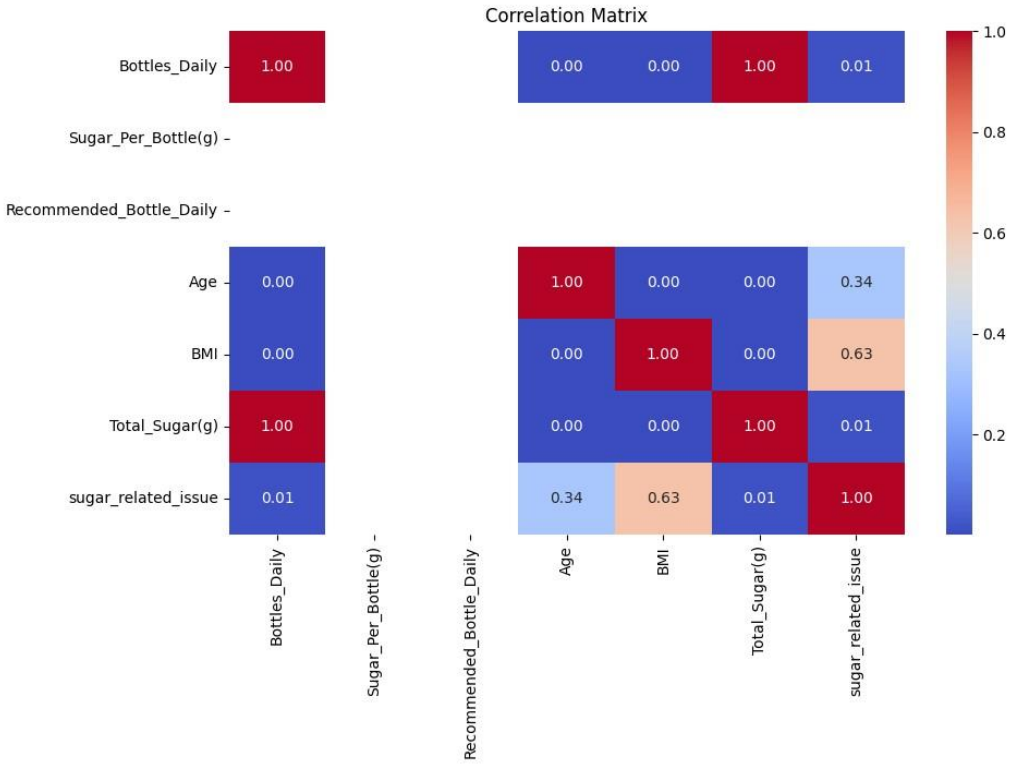


Fig. 8(b) Correlation Matrix

In this research, investigating the impact of soft drink consumption on human health, a correlation matrix serves as a crucial tool to explore the relationships between different variables. Each cell in the correlation matrix represents the correlation coefficient between two variables, indicating the strength and direction of their linear relationship. Analyzing this matrix helps us gain insights into how various factors associated with soft drink consumption, such as sugar content, volume consumed, frequency of consumption, and individual health indicators, are interrelated. For instance, positive correlations between soft drink consumption and health-related issues like obesity, diabetes, or dental problems suggests potential adverse effects of excessive soft drink intake on human health.

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

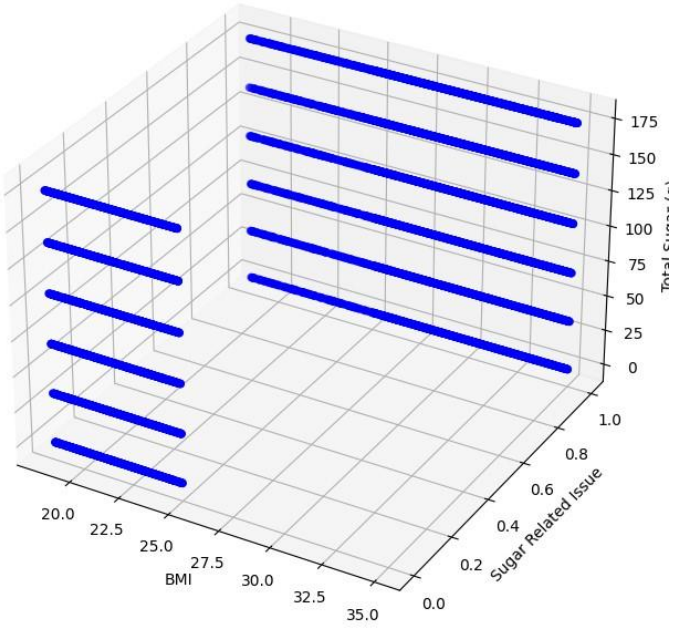


Fig.9 Scatter Plot of BMI, Sugar Related Issue and TotalSugar

BMI, sugar-related issues, and total sugar intake are vital factors in assessing the impact of soft drinks on human health, offering lots of insights into both individual and population-level health outcomes. BMI serves as a fundamental indicator of weight status, in this research, we used it to assess the association between soft drink consumption and obesity. High consumption of sugary beverages has been linked to weight gain and obesity, as these drinks often contribute to excess calorie intake and do not promote satiety. Consequently, monitoring BMI trends alongside soft drink consumption helps identify individuals and populations at increased risk of obesity-related health complications, such as cardiovascular disease, type 2 diabetes, and certain cancers. By integrating BMI data into this research data for health assessments, we evaluate the effectiveness of interventions aimed at reducing soft drink consumption and mitigating the burden of obesity-related diseases.

Similarly, sugar-related issues encompass a range of health conditions associated with excessive sugar intake, including dental caries, metabolic syndrome, and insulin resistance. By examining the prevalence of sugar-related issues among individuals who consume high quantities of soft drinks, we elucidate the direct impact of these beverages on metabolic health and overall well-being.

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

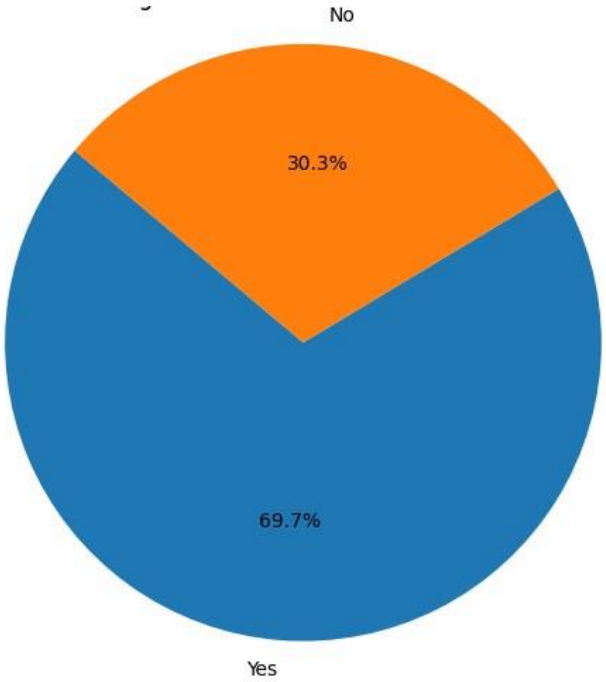


Fig.10 A pie chart showing the Distribution of Risk (Class)

The pie chart in Figure 10 illustrates the distribution of the risk of developing sugar-related health issues among individuals based on their soft drink consumption habits. Notably, 69.7% of individuals are identified as at risk of experiencing sugar-related health complications due to consuming high volumes of soft drinks daily. Conversely, 20.3% of individuals did not exceed the threshold volume associated with risking such health complications. This visualization underscores the significant proportion of individuals susceptible to sugar-related health issues due to their soft drink consumption patterns, highlighting the importance of promoting healthier beverage choices and implementing targeted interventions to mitigate associated health risks.

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

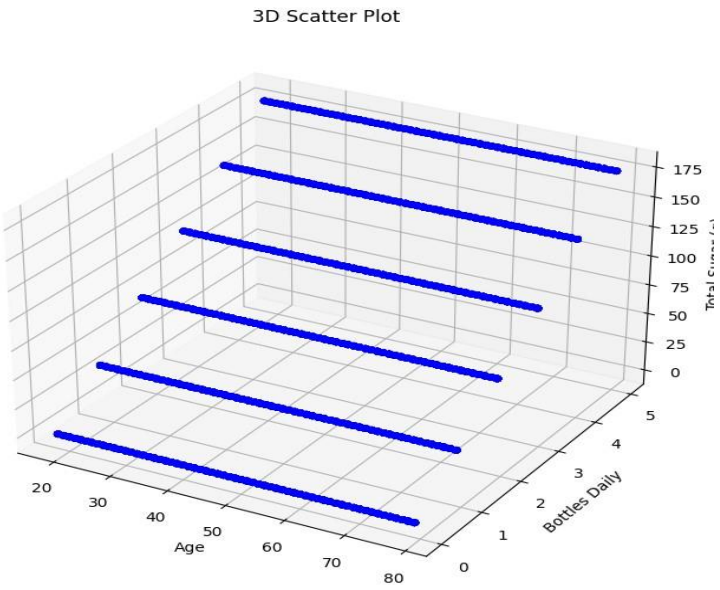


Fig.11 Scatter Plot of Age, Bottles_Daily and TotalSugar

The contribution of age, daily soft drink consumption (Bottles_Daily), and total sugar intake to the risk of developing sugar-related health complications due to soft drink consumption is important and pivotal in understanding the complex interplay between these factors and individual health outcomes. Age serves as a critical determinant, as older individuals may be more susceptible to the adverse effects of excessive sugar intake due to age-related changes in metabolism and physiological resilience. Additionally, age-related factors such as declining bone density and compromised dental health further heighten the risk of developing sugarrelated health complications, including osteoporosis and dental caries. Furthermore, older individuals may have accumulated years of soft drink consumption, leading to a higher cumulative exposure to dietary sugars and subsequent health risks. Thus, age emerges as a significant contributor to the risk of sugar-related health complications, revealing the importance of considering age-specific interventions and health promotion strategies to mitigate associated risks among vulnerable populations. Moreover, the frequency and volume of daily soft drink consumption (Bottles_Daily) play a pivotal role in determining an individual's risk of sugar-related health complications. High levels of soft drink consumption are directly associated with increased sugar intake, contributing to excess calorie consumption and a higher likelihood of developing obesity, type 2 diabetes, and cardiovascular diseases. Individuals who consume multiple bottles of soft drinks daily are at elevated risk due to the cumulative effect of prolonged exposure to dietary sugars. Additionally, the total sugar content in soft drinks directly contributes to the overall sugar intake and subsequent health risks, with higher sugar content correlating with increased risk of developing sugar-related health complications. Therefore,

ISSN: 3064-8270

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

understanding the combined impact of age, daily soft drink consumption, and total sugar intake provides valuable insights into the multifactorial nature of sugar-related health risks and informs targeted interventions to promote healthier dietary habits and reduce the burden of associated health complications.

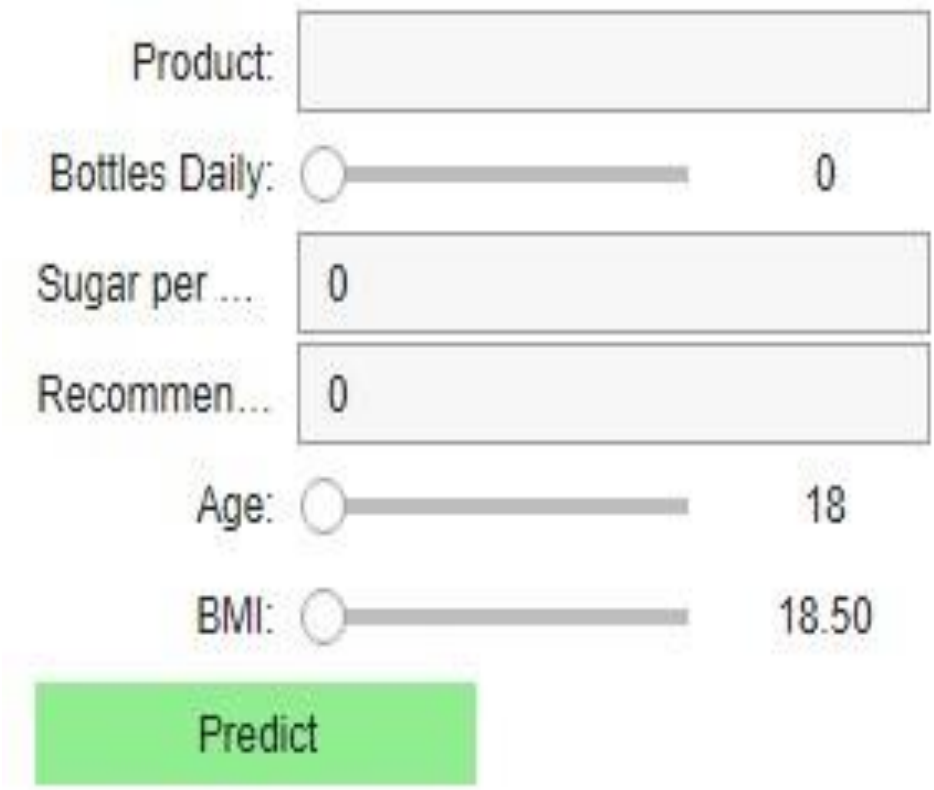


Fig. 12 API Deployment

In the context of this research on assessing the health impact of soft drink consumption, the development of an interface or API for the model holds significant relevance and practical importance. Such an interface would serve as a gateway for stakeholders, including healthcare professionals, policymakers, and individuals, to access and utilize the model's predictions and insights in a user-friendly manner. For instance, healthcare professionals could integrate the model's predictions into electronic health record systems or clinical decision support tools to assess patients' risk of sugar-related health complications based on their soft drink consumption habits. Policymakers could utilize the model's insights to inform public health initiatives and interventions aimed at promoting healthier dietary habits and reducing the prevalence of soft drink consumption. Moreover, an interface or API for the model would enhance its accessibility and scalability, enabling deployment across

various platforms and environments. For instance, the model could be deployed as a web service accessible through standard HTTP requests, allowing individuals to access its predictions through web browsers or mobile applications. Additionally, the interface could support real-time processing of data streams, enabling continuous monitoring of soft drink consumption patterns and health outcomes at population levels. By providing a standardized interface for interaction, the model becomes more versatile and adaptable to diverse use cases and environments, fostering broader adoption and impact in addressing the public health concerns associated with excessive soft drink consumption.

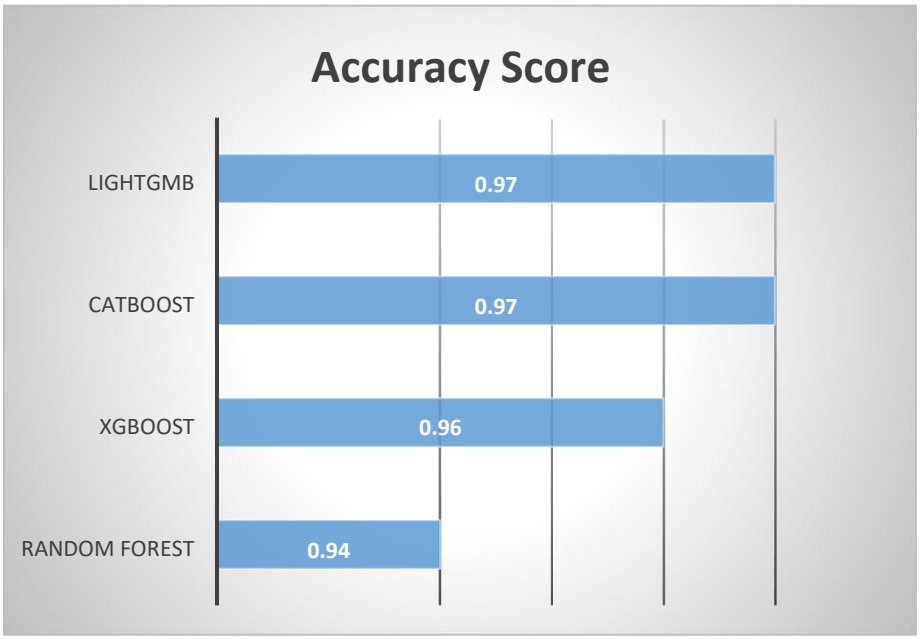


Fig.13 Comparative Analysis of the Performance of the four Ensemble Learning Techniques In our machine learning model, we employed ensemble learning techniques, using advanced algorithms such as LightGBM, CatBoost, XGBoost, and Random Forest. Notably, LightGBM and CatBoost emerged as the top-performing models, yielding an impressive accuracy score of 97%. These models demonstrated remarkable predictive capabilities, showcasing their efficacy in accurately classifying instances and capturing complex patterns within the data. Following closely behind, XGBoost achieved a commendable accuracy score of 96%, further solidifying its position as a robust algorithm for classification tasks. Lastly, although Random Forest exhibited slightly lower accuracy compared to the other models, it still delivered a respectable performance with an accuracy score of 94%. Overall, the ensemble learning framework allowed us to harness the strengths of diverse algorithms, enabling us to build a robust predictive model capable of effectively addressing the complexities of our dataset.

5.0 CONCLUSION

The consumption of soft drinks, particularly those high in added sugars, has become a prevalent habit globally, raising concerns about their potential health impacts. These beverages have been associated with various adverse health effects, including obesity, type 2 diabetes, cardiovascular diseases, and dental problems. Given the rising public health concerns, innovative approaches are crucial to assess their impact accurately. In this study, we employed four Ensemble Learning approaches to investigate the health impact of soft drink consumption.

Our findings reveal that older individuals may not require as much soft drink consumption, and the general population should limit their daily intake to a single bottle, approximately 35g for a 35cl content. This research aims to identify the health risks associated with exceeding the recommended consumption of 35cl Cola products in a day.Determining the permissible quantity of Cola consumption without encountering sugar-related issues depends on various factors, including individual health, activity level, overall diet, and existing health conditions such as diabetes or obesity. However, health organizations such as the World Health Organization (WHO) recommend limiting added sugar intake to no more than 10% of total daily calories, which translates to approximately 50 grams per day for an average adult. Considering that a 330ml bottle of Coca-Cola Classic contains around 35 grams of sugar, consuming one bottle already exceeds a significant portion of the recommended daily sugar intake. Consuming multiple bottles in a day could easily surpass the recommended limit. Therefore, it's advisable to restrict the consumption of sugary beverages like Cola and opt for healthier alternatives such as water, unsweetened tea, or sparkling water flavored with natural ingredients to mitigate the risk of adverse health outcomes associated with excessive sugar consumption.

REFERENCES

Doris, O. S., Ikechukwu, P. and Ejidike, S.A. (2019). HPLC determination of benzoic acid, saccharin, and caffeine in carbonated soft drinks. *International Journal of ChemTech Research*, 12(4), 15-23. https://doi.org/10.20902/IJCTR.2019.120403.

Azuma, S. L., Quartey, N. K. A. and Ofosu, I. W. (2020). Sodium Benzoate in Non-alcoholic Carbonated (soft) Drinks: Exposure and health risks. *Scientific African*, 10, e00611.

https://doi.org/10.1016/ j.sciaf.2020.e00611.

Esrafil, M., Akter, S., Alam, M. J., Haque, M. A., Zubair, M. A. and Khan, M.S. H (2022). Identification and quantification of sodium benzoate in soft drinks available in Tangail region by high-performance liquid chromatography. Food Research 6 (3) : 220 - 225.

Joseph, S. A., Patrice, A. C. O., Daniel, O. (2022). Effect of storage on the levels of sodium benzoate in soft drinks sold in some Nigerian market with exposure and health risk assessment. Environmental Analysis Health and Toxicology, 37(4),pp.1-10.

Akolawole, J. S., Patrice, A. C. O., Daniel, O. (2022). Effect of storage on the levels of sodium benzoate in soft drinks sold in some Nigerian market with exposure and health risk assessment. Journal of Environmental Analysis health and Technology, 37(4),1-10.

Esrafil, M., Akter, S., Alam, M.J., Haque, M. A., Zubair, M. A. and Khan, M. S. H (2022). Identification and quantification of sodium benzoate in soft drinks available in Tangail region by high-performance liquid chromatography. Journal of Food Research 6 (3), 220 - 225.

Tahmassebi, J. F., & BaniHani, A. (2019). Impact of soft drinks to health and economy: a critical review. Journal of Public Health, 27(6), 677-682. doi: 10.1007/s10389-01901024-7

BaniHani, A., & Tahmassebi, J. F. (2019). Impact of soft drinks to health and economy: a critical review. European Archives of Paediatric Dentistry, 20(3), 175-182. doi: 10.1007/s40368-019-00458-0

Anthony Edet, Uduakobong Udonna, Immaculata Attih, and Anietie Uwah (2024). Security Framework for Detection of Denial of Service (DoS) Attack on Virtual Private Networks for Efficient Data Transmission. Research Journal of Pure Science and Technology, 7(1),71-81. DOI: 10.56201/rjpst.v7.no1.2024.pg71.81

Philipp K., Sven S., Jutta M. (2016). Soft drink consumption and mental health problems: Longitudinal relations in children and adolescents. Canadian Journal of Public Health, 107(3), e224-e229. doi: 10.17269/cjph.107.5325

Harvard T. H. Chan School of Public Health. (2023). Sugary Drinks. Retrieved from https://www.hsph.harvard.edu/nutritionsource/healthy-drinks/sugary-drinks/

Vartanian, L. R., Schwartz, M. B., & Brownell, K. D. (2007). Effects of soft drink consumption on nutrition and health: a systematic review and meta-analysis. American Journal of Public Health, 97(4), 667-675. doi: 10.2105/AJPH.2005.083782

De Koning, L., Malik, V. S., Kellogg, M. D., Rimm, E. B., Willett, W. C., & Hu, F. B. (2012). Sweetened beverage consumption, incident coronary heart disease, and biomarkers of risk in men. Circulation, 125(14), 1735-1741. doi: 10.1161/CIRCULATIONAHA.111.067017

Hu H, Song J, MacGregor GA, He FJ. Consumption of Soft Drinks and Overweight and Obesity Among Adolescents in 107 Countries and Regions. JAMA Netw Open. 2023;6(7):e2325158. doi:10.1001/jamanetworkopen.2023.25158

Brown, R. J., & Rother, K. I. (2023). Artificial sweeteners: a systematic review of metabolic effects in humans. Diabetes, Obesity and Metabolism, 25(1), 3-8. doi: 10.1111/dom.14500

Popkin, B. M., Hawkes, C., & Sweeting, A. N. (2023). Soft drink consumption: a global perspective. Nutrition Reviews, 81(1), 1-14. doi: 10.1093/nutrit/nuz080

Rubaiya H., Mohammad R. H., Aniruddha R.t, Mohammad S. U. (2022). Image-based soft drink type classification and dietary assessment system using deep convolutional neural network with transfer learning. Journal of King Saud University-Computer and Information Sciences, 34(5), 101487. doi: 10.1016/j.jksuci.2020.08.015

Edet, A. E. and Ansa, G. O. (2023). Machine learning enabled system for intelligent classification of host-based intrusion severity. Global Journal of Engineering and Technology Advances,16(03), 041–050.

Ekong, B., Ekong, O., Silas, A., Edet, A., & William, B. (2023). Machine Learning Approach for Classification of Sickle Cell Anemia in Teenagers Based on Bayesian Network.

Journal of Information Systems and Informatics, 5(4), 1793-1808. https://doi.org/10.51519/journalisi.v5i4.629.

Ekong, A., Ekong, B., and Edet, A. (2022), Supervised Machine Learning Model for Effective Classification of Patients with Covid-19 Symptoms Based on Bayesian Belief Network, Researchers Journal of Science and Technology(2022),2, pp-27-33.

Uwah, A. and Edet, A. (2024).Customized Web Application for Addressing Language ModelMisalignment through Reinforcement Learning from HumanFeedback. World Journal of Innovation And Modern Technology,8,(1), 62-71. DOI: 10.56201/wjimt.v8.no1.2024.pg62.71.

I. J Umoren & S. J. Inyang, "Methodical Performance Modelling of Mobile Broadband Networks with Soft Computing Model," International Journal of Computer Applications, vol. 174, no. 25, pp. 7-21, 2021.

S. Inyang and I. Umoren (2023) "From Text to Insights: NLP-Driven Classification of Infectious Diseases Based on Ecological Risk Factors," Journal of Innovation Information Technology and Application (JINITA), vol. 5, no. 2, pp. 154-165,

S. Inyang and I. Umoren (2023) "Semantic-Based Natural Language Processing for Classification of Infectious Diseases Based on Ecological Factors,". International Journal of Innovative Research in Sciences and Engineering