The Relation Among Obesity and Sugar Consumption: A Machine Learning Approach

Nobonita Saha¹, Aninda Mohanta², Jannatun Tuba Jyoti³, Diti Roy⁴ and Tamal Joyti Roy⁴

¹Institute of Nutrition and Food Science, University of Dhaka

²Department of Computer Science and Engineering
Institute of Engineering and Management, Kolkata, India

³Department of Computer Science and Engineering
North Western University, Khulna

⁴Institute of Information and Communication Technology
Khulna University of Engineering & Technology (KUET)

nobonitasahaqueen@gmail.com, anindamohanta21@gmail.com,
jannatuntuba@gmail.com, diti267@gmail.com, tjroy13june@gmail.com

Abstract—Obesity is now a common problem for the people of all over the world. People of all ages are suffering from obesity and that's why it has become a real public health problem to us. There are many reasons for obesity like high consumption of fat, sugar foods, eating disorders, lack of physical activity, genetic etiology and many more. In our study, we have tried to show the sugar association result with obesity which is performed by machine learning algorithms. We have used two data sets of the USA to complete this study. We have preprocessed the data by different techniques like feature selection, data cleaning, and data transforming. Then we have applied nine machine learning algorithms for both data set. For the first data set, We have found the best accuracy from Naïve Bayes and Random Forest which is between 74-98%. Also, we have found the best accuracy for the second data set from KStar and Naïve Bayes which is between 78-97%. Our study is also able to show that not only the old aged people but also the young generation are suffering from obesity because of sugar.

Index Terms—Obesity, USA, Digital-Health, Clinical-Health, Random Forest, Decision Table, Classification

I. INTRODUCTION

Over the last decade, the speedy upward prevalence of Obesity has emerged as one of the distinguished issues in present global public health science [1], [2]. Rendering to WHO, cumbersome and fatness are defined as irregular or unnecessary fat buildup which may damage fitness [3]. It is a state of excess body weight due to fat deposition compared with the set standards of body weight. These excess weights may come from muscles (lean body mass), bone, fat (adipose tissues), sometimes from tumors and/or body water (edema). But obesity especially refers to having an abnormally high proportion of total body fat (WHO, 1998; NHLBI, 1998). The main etiology or the cause of obesity is the imbalance between energy expenditure and energy consumption [4]. Moreover, environmental changes, lifestyle and behavioral changes, diet modification, unhealthy habits (such as consumption of excess alcohol and tobacco), physical inactivity, eating disorders, individual variables (age, sex, ethnicity, etc), and genetic influences promote the development of obesity [5]. Corpulence and its related long-lasting sicknesses are noteworthy community

health matters in the worldwide situation with considerable costs to both individuals and the community. Obese individuals have many long and short-term negative health consequences [6]. Obesity is one of the major contributors to the development of hypertension, cardiovascular diseases, elevated plasma insulin concentrations and insulin resistance, diabetes mellitus and hyperlipidemia, sleep-breathing abnormalities, gallbladder diseases, osteoarthritis, and some form of cancers [7]. 'Conferring to the WHO, worldwide plumpness has closely triplicated later 1975. In 2016, more than 1.9 billion adults were overweight aged less than or equal to 18 years (among them 650 million were obese), 340 million children were overweight or obese aged 5-16 years. In recent 2020, about 39 million children under the age of 5 were overweight or obese. The total number of obese people in the world is about 2.1 billion, accounting for about 30% of the total population. And the incidence prevalence rate continues to rise upward. Each year over 3 million people die from obesity. So, it is clear that the consequent rise of obesity has the potentials to increase the risk of mortality and morbidity of individuals. The burden of obesity has been steadily rising in low and middle-income countries and the trend is not significantly different in most of the countries. If the trend continues to rise, it is a close estimate that one-third of the adult population will become overweight, and about more than 1 billion people will be obese worldwide by 2025 [8]. While we approach obesity-related problems, machine learning algorithms can be used for studying the link between sugar consumption and obesity. The primary purpose of this paper is not to identify all risk predictions of obesity rather than the main aim is to carry out the application, analyzing and evaluation of machine learning algorithms in assessing and identifying obesity risk factors and their associations using sugar factors as inputs with high BMI (obese) to machine learning. Our mechanisms add insights to important clinical experiments at individual levels. The application of machine language shows that models can achieve significantly better results.

II. RELATED WORKS

Obesity is a complex, multifactorial, and major public health problem worldwide that involves an excessive amount of body fat which increases the risk of developing diseases and other health difficulties, such as heart disease, diabetes, high blood pressure, and certain types of cancer [10]. Decision Tree (DT), Support vector machine (SVM), The Gaussian Naive Bayes (GNB) classification, Logistic regression (LR), Random Tree (RT), J48, ID3, Naïve Bayes, BayesNet and so on for any predictive analysis [11]. In this paper, T. M. Dugan, S. Mukhopadhyay, A. Carroll and S. Downs, the author of Ref. [12] tried to develop a machine learning model to predict childhood obesity. With this desire, they used six different machine learning methods. Casimiro Aday Curbelo Montañez et al. the author of [13] and the performances were evaluated using the Area Under the Curve (AUC), sensitivity (SE), and specificity (SP). Finally, they obtained 90.5% accuracy with the help of a support vector machine. Balbir Singh and Hissam Tawfik the author of [14] was interested to construct a machine learning-driven model to predict the risk of childhood obesity at the age of 14. Rodolfo Cañas Cervantes and Ubaldo Martinez Palacio the author of created an intelligent method [15] based on supervised and unsupervised techniques of data mining to detect obesity levels. For this purpose. Casimiro Aday Curbelo Montañez et al, the author of presented a genetic profile [16] predictive study using a Random Forest-based feature selection algorithm. And it was revealed as Support Vector Machine attained the best predictive performance among their studied models. And they obtained 90.5% accuracy. In this study a economic crisis during COVID-19 had shown with the help of machine learning algorithm [17] they used 10-fold cross validation for the experiment. Accuracy were different with different algorithms. This experiment showed a future and ethics of artificial intelligence with health care [18], where a lot of areas of machine learning and artificial intelligence had been discussed.

III. PROPOSED METHODOLOGY

We have collected two data sets, [19], [20]. Figure1 showing the overall working procedure. First of all, we collected the datasets each of which had a separated shape. For example, the data set 1 shape was (43338,6) and the data set 2 shape was (45000,9). We preprocessed it with feature selection, data cleaning, and transform the data. 1325 missing values mean either the data cells were empty or filled with lexical symbols. After the preprocessing, we split each data set. There were 70% training data and 30% test data. The sample data sets are showing in Tables 1 and 2. Algorithm 1 is our proposed algorithm for finding the desired data sets. Where the D_n to D_{n+1} connected with the data set 1 and data set 2. We used a sorting algorithm for sorting our data values. There were missing values that were later demolished in the results. No garbage values were considered. Figure 2 showing the MLP algorithm's visualization, the input nodes and output nodes were divided with a hidden layer named Bias and the other

denoted with H1 and H2, after calculating and compare the hidden layers turn into like this ($H1 : 1$) and ($H1 : 2$).

TABLE I
SAMPLE DATA SET 1

fat 100g	Carbohydrates 100g	sugars 100g	proteins 100g	salt 100g	energy 100g	reconstructed energy	g_sum	product Name	Sugar Factor
28.57	64.29	14.29	3.57	0	2243	2267.85	96.43	Banana Chips Sweetened (Whole)	0
17.86	60.71	17.86	17.86	0.635	1941	2032.23	96.43	Peanuts	0
57.14	17.86	3.57	17.86	1.22428	2540	2835.7	92.86	Organic Salted Nut Mix	0
18.75	57.81	15.62	14.06	0.1397	1833	1953.04	90.62	Organic Muesli	0
36.67	36.67	3.33	16.67	1.60782	2230	2336.91	90.01	Zen Parry Mix	0
18.18	60	21.82	14.55	0.02286	1824	1976.37	92.73	Cinnamon Nut Granola	0

TABLE II
SAMPLE DATA SET 2

PCT DIABETES ADULTS09	PCT DIABETES ADULTS10	PCT OBESE ADULTS09	PCT OBESE ADULTS10	PCT OBESE ADULTS13	PCT OBESE CHILD08	PCT OBESE CHILD11
12.3	11.8	34.1	30.5	32.4	12.4	13.5
11.4	11.8	25.5	26.6	32.4	10.8	13.3
14	14.2	37	37.3	32.4	20.8	18.6
11.8	11.1	34	34.3	32.4	15.4	13.7
12.6	14	32	30.4	32.4	19.9	18.5

For our experimental study, for data set 1, we have calculated a standard value of daily sugar consumption. We have made a class that holds the values of the sugar percentage of each food where, $sugar > 25gram$). We found a significant amount of products that were more in sugar, even more than our daily needs. There were 10820 products, using more than 25gram of sugar in their products. To test the accuracy of our findings we used nine machine learning classifiers algorithms. The first data set accuracy for its defined class ranges between 74-98%. The first data validation's completion we tested the second data set which consisted of the obesity rate of the USA. A healthy human body mass index varies between 22-25.6 or something like that(it varies in age, gender) that's why we took the BMI values from age 10-50 and 65-80 and check the data validity rate. Processed those validations with logic like, if the BMI is greater than 26 then we consider that individual unhealthy. This way we have created another class and tested it with our nine machine learning algorithms. The accuracy rate of each algorithm varies between 78-97%. Tables 2 and 3 showing the test data set machine learning results, where the whole data converted into an array to perform the analysis. We have used the jupyter notebook for our experimental study.

Algorithm 1 DS1,DS2,SFact,OFact

- 1: start
 - 2: Read DS1 and DS2
 - 3: Compare D_n to D_{n+1}
 - 4: The middle value calculated of M_{D+1} to $D_n = d + (d-1)/2$
 - 5: Ratio $D_{rate} = D_n, D_{n+1}, m, n$
 - 6: Result $D_n = N, D_n, n$
 - 7: End
-

TABLE III
CORRELATION OF DATA SET 1

	CC	CC2	Covariance
sugars_100g	1	.866**	462.057
SugarFactor	866**	7.94	7.949
N			45028

Fig. 1. Overall Flow Chart

IV. EXPERIMENTAL RESULTS

TABLE IV
CORRELATION OF DATA SET 2

	CC	CC2	Covariance
PCT DIABETES ADULTS09	.952**	.148**	.100**
PCT DIABETES ADULTS10	866**	7.94	7.949
PCT OBESE CHILD11	.669**	.103**	45028
PCT OBESE CHILD08	7.94	.952**	
N	3138	3138	2714

The nine machine learning algorithms' accuracy for both data set showing in Figures 2 and 3. The best accuracy gained for data set 1 was the Naïve Bayes, Random Forest and for data set 2, the best-performed algorithms were KStar and Naïve Bayes. So there had been a sign, a relation that sugar is impacting our health. Can have a direct connection to the Obesity rate in the USA. Table 4 showed the significant correlation among the data. The Sugar Factor is directly correlated with sugar consumption (which is per 100 gram). Our finding of the Correlation of data set 2. The adult in

two age range and child in two age range considered for the correlation analysis. There were three input layers, three hidden layers, and one output layer and that is the class "Sugar_100g". It indicated that the sugar quantity in each product that we have collected showed a significant amount of excessive results. That means there was more sugar in every product than there were raw ingredients. Our machine learning prediction results showed that the products which were excess of sugar, perfectly fit in our derived classes. Furthermore, data set 2 also correlated with the diabetes patient's sugar intake. It also compared with the Obsessed but not diabetic individuals. One interesting thing for our findings in experimental results, that is there were more children linked to obesity, and their daily sugar consumption also excessive. The ratio of sugar intake and children's consumption was 0.866 and the factor's covariance was 7.949. The age above 60, individuals data set and their factors showed almost 0.97 ratios inaccuracy, that means those who are older than 60, consuming more than 30gram of sugar on daily basis were obese and diabetic. Our machine learning approaches in data set 1 and 2 showed the same results that were illustrated in figure 2 and figure 3. The bayesian's worked always well.

Fig. 2. Accuracy of different machine learning algorithms for data set 1

Fig. 3. Accuracy of different machine learning algorithms for data set 2

V. CONCLUSION

Nowadays, obesity is a common term for us and it seems that almost all types of aged people are suffering from this problem. It is so much harm to our health. That's why we have tried to find out the reason for this problem which is occurred by sugar consumption and finally, we can show the

result by the machine learning algorithm. We have taken two data sets for four types of aged people like 0-8,8-10,0-50 and 50-60 and applied nine machine learning algorithms for both data set. We have got the best accuracy for both data set. Last of all it is founded that eating more sugar is very bad for our health and it is one of the causes of obesity. Also, it is founded that from our study, all types of aged people are suffering from this problem. Our study only shows the obesity factor because of excess sugar intake but in the future, we will try to show other reasons for obesity.

REFERENCES

- [1] R. J. R. Levesque, "Obesity and Overweight," *Encycl. Adolesc.*, pp. 1–5, 2017, doi: 10.1007/978-3-319-32132-5_447-2.
- [2] X. Cheng et al., "Does Physical Activity Predict Obesity—A Machine Learning and Statistical Method-Based Analysis," *Int. J. Environ. Res. Public Heal.* 2021, Vol. 18, Page 3966, vol. 18, no. 8, p. 3966, Apr. 2021, doi: 10.3390/IJERPH18083966.
- [3] "Obesity and overweight," <https://www.who.int/news-room/factsheets/detail/obesity-and-overweight> (accessed Jul. 14, 2021).
- [4] L. Serra-Majem and I. Bautista-Castaño, "Etiology of obesity: two 'key issues' and other emerging factors," *Nutr. Hosp.*, vol. 28, no. 5, pp. 32–43, 2013, Accessed: Jul. 14, 2021. [Online]. Available: <https://www.redalyc.org/articulo.oa?id=309229028004>.
- [5] A. Khan A. and A. Khan, "Prevalence and Etiology of Obesity - An Overview," *Pakistan J. Nutr.*, vol. 3, no. 4, pp. 14–25, 2004, doi: 10.1.1.323.5274.
- [6] P. G. Kopelman, "Obesity as a medical problem," *Nat.* 2000 4046778, vol. 404, no. 6778, pp. 635–643, Apr. 2000, doi: 10.1038/35007508.
- [7] G. A. Bray, "Medical Consequences of Obesity," *J. Clin. Endocrinol. Metab.*, vol. 89, no. 6, pp. 2583–2589, Jun. 2004, doi: 10.1210/JC.2004-0535.
- [8] T. Omer, "The causes of obesity: an in-depth review," *Adv. Obesity, Weight Manag. Control*, vol. 10, no. 4, pp. 90–94, Jul. 2020, doi: 10.15406/AOWMC.2020.10.00312.
- [9] H. Çakmur, "Obesity as a Growing Public Health Problem," *Adiposity - Epidemiol. Treat. Modalities*, Mar. 2017, doi: 10.5772/65718.
- [10] "Obesity - Symptoms and causes - Mayo Clinic," <https://www.mayoclinic.org/diseases-conditions/obesity/symptoms-causes/syc-20375742> (accessed Jul. 14, 2021).
- [11] O. Babajide et al., "A Machine Learning Approach to Short-Term Body Weight Prediction in a Dietary Intervention Program," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12140 LNCS, pp. 441–455, Jun. 2020, doi: 10.1007/978-3-030-50423-6_33.
- [12] "Europe PMC," <https://europepmc.org/article/pmc/pmc4586339> (accessed Jul. 14, 2021).
- [13] C. A. C. Montanez et al., "Machine learning approaches for the prediction of obesity using publicly available genetic profiles," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, pp. 2743–2750, Jun. 2017, doi: 10.1109/IJCNN.2017.7966194.
- [14] B. Singh and H. Tawfik, "Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People," *Comput. Sci. – ICCS 2020*, vol. 12140, p. 523, 2020, doi: 10.1007/978-3-030-50423-6_39.
- [15] R. C. Cervantes and U. M. Palacio, "Estimation of obesity levels based on computational intelligence," *Informatics Med. Unlocked*, vol. 21, p. 100472, Jan. 2020, doi: 10.1016/J.IMU.2020.100472.
- [16] P. F. Casimiro Aday Curbelo Montañez, B. Abir Hussain, Dhiya Al-Jumeily, and J. H. and N. R. Abdulaimma, "NoMachine Learning Approaches for the Prediction of Obesity using Publicly Available Genetic Profiles Title."
- [17] D. Roy, T. J. Roy, and M. A. Mahmood, "An Efficient Approach to Identify Economic Crisis During Covid-19 Outbreaks Utilizing Data Mining," *SSRN Electron. J.*, May 2021, doi: 10.2139/SSRN.3852813.
- [18] S. Bhattacharya, M. M. Hossain, R. Juyal, N. Sharma, K. B. Pradhan, and A. Singh, "Role of public health ethics for responsible use of artificial intelligence technologies," *Indian J. Community Med.*, vol. 46, no. 2, pp. 178–181, Apr. 2021, doi: 10.4103/IJCM.IJCM_62_20.

- [19] “Fast Food Chains - Nutrition Values — Kaggle.”
<https://www.kaggle.com/denizyigit/fast-food-chains-nutrition-values>
(accessed Jul. 10, 2021)
- [20] “adamhelsinger/food-environment-atlas — Workspace — data.world.”
<https://data.world/adamhelsinger/food-environment-atlas/workspace/file?filename=DataDownload.xls> (accessed Jul. 10, 2021).