Research Article

Machine Learning-Based Clustering and Classification of Obesity Associated with Elevated Screen Time

Muhammad Ammar [1^], Abdur Rauf [12^], Nasreen Noor [1], Mahnoor Azhar [1], and Syeda Marriam Bakhtiar [1*]

[1] Genetic and Molecular Epidemiology Research Group, Department of Bioinformatics and Biosciences, Capital University of Science and Technology, Islamabad, Pakistan; mammarhrn@gmail.com (M.A.); abdurrauf99@gmail.com (A.R.); nisreennoor2017@gmail.com (N.N.); mahnoorazhar88@gmail.com (M.A.); smarriamb@gmail.com (S.M.B.);
[2] National University of Science and Technology, Islamabad, Pakistan; abdurrauf99@gmail.com (A.R.);
* Correspondence: smarriamb@gmail.com (S.M.B.);
^ Share Equal Contribution

Abstract: Obesity is a complex chronic disease characterized by excessive fat deposits that can undermine health. Obesity can lead to increased risk of type 2 diabetes and cardiovascular diseases; it can affect bone health and reproduction; and it also increases the risk of certain cancers. WHO in 2022 reported 2.5 billion people are suffering from obesity. It is a multifactorial disease due to obesogenic environments, psychosocial factors, and genetic predispositions. Lacks of physical activity, sedentary lifestyle, mode of transportation, and screen time have also shown significant impact on it. Because of the strong link between elevated screen time and obesity, effective and advanced studies are required to address and understand this issue. Classification, prediction, and population-level analysis help us to build future prevention strategies to combat. Machine learning algorithms proved their remarkable results in the latest studies for prediction analysis. We have used data from the open source library UCI and applied three machine learning approaches for classification, prediction, and cluster analysis: K-mean clustering, the K-nearest neighbor algorithm (KNN), and a decision tree with a random forest. Among them all, the decision tree with random forest has shown the best results with an accuracy of 95.64% and a relative absolute error of 17.96%. All approaches proved strong associations between increased screen time and obesity level.

Keywords: Machine Learning; Obesity; Screen Time; Association; Public Health; Decision Tree; K–Nearest Neighbor; Random Forest; K-Mean Clustering;

Citation: Ammar, M.; Rauf, A.; Noor, N.; Azhar, M.; Bakhtiar, S. M. Machine Learning-Based Clustering and Classification of Obesity Associated with Elevated Screen Time. *Pakistan Journal of Biochemistry and Biotechnology*, **2024**, 5 (2), 2024. https://doi.org/10.52700/pjbb.v5i2.241
Received: date: 4 April 2024
Accepted: date: 22 October 2024
Published: date: 30 October 2024

1. Introduction

The phrase "screen time" describes the amount of time people spend using electronic devices, and its ubiquity is linked to the gadgets' widespread use worldwide [1]. With 66.7% of people using mobile phones worldwide and 86% of households in Pakistan owning one, screen-based gadgets are now readily available sources of entertainment, information, and advertising. Several studies show a direct link between obesity and screen time because obesity is associated with a higher Body Mass Index (BMI), which has a detrimental effect on psychological well-being.

Screen time's sedentary nature increases the likelihood of physical inactivity, which throws off regular hunger signals and encourages the ingestion of foods that increase body weight. Although a lot of screen time is a sign of a sedentary lifestyle, screen time alone does not directly promote weight gain [2,3]. Examine the integration of SVM into the KNN classifier in the context of machine learning.

There is rising concern about how screen time affects children's health, especially about obesity, and some associations are calling for limits on children's screen time [4]. Notwithstanding, certain research findings have led to

criticism of these guidelines, with some advocating for a more nuanced approach that takes into account the context and content of screen usage. Maintaining a healthy weight requires physical activity; thus, it's alarming to see the drop in physical activity, particularly among youth [5].

There are fewer possibilities for in-school physical activity since students of color in impoverished schools have difficulty receiving physical education. The "couch-potato" hypothesis was put forth by research that connected the use of electronic media to obesity, albeit the exact causation pathways are still unknown. Increased screen time has been linked to inadequate sleep, which has been linked to poor health outcomes in children and adolescents, including obesity [3,6].

Throughout industries, including e-commerce and healthcare, machine learning (ML) techniques are being used more and more to evaluate massive datasets and offer customized solutions [7]. Machine learning techniques (SVMs, DT, and KNN) are used to process complex data in dynamic situation, regression, and classification. In binary classification problems, SVMs perform very well, KNN clusters data points according to their closeness, and decision trees are hierarchical in structure. As per recent studies, a machine learning technique provides tailored insight into data and the best solutions [8].

Numerous studies have demonstrated the ability of machine learning to uncover important variables and enhance intervention tactics, making it a formidable tool in the prediction of obesity. A noteworthy study that employed LASSO regression and supervised machine learning to identify both conventional and unique risk factors impacting BMI percentiles was conducted on kids who identified as Hispanic or Latino [19]. The study demonstrated that behavioral and psychological factors, in addition to traditional measures like food and physical activity, play important roles in obesity and underscored the need for customized obesity interventions that take into account the distinct sociocultural settings of various racial and ethnic groups.

Utilizing easily accessible data to build prediction models, another important contribution to the subject made use of electronic health records to predict childhood obesity. The study showed that some metrics were important for predicting obesity by age five, including weight-for-length z-scores and BMI readings obtained before the age of two. With area under the curve values of 81.7% for girls and 76.1% for boys, the machine learning algorithms used to create the system demonstrated excellent prediction accuracy and suggested a strong possibility for early detection of children who are at risk [20,21]. These results demonstrate the revolutionary potential of machine learning in obesity prediction.

2. Materials and Methods

2.1. K-Means Clustering Algorithm

In the K-means clustering algorithm, data points are grouped into clusters based on how close they are to centroids. It is a basic machine learning technique. Researchers use K-means clustering to identify cognitive patterns connected with certain health issues [9]. The algorithms' mathematical foundation is demonstrated through equations [10]. Equation 1 and equation 2 show how to calculate the formula for Euclidean distance and the new centroid value, respectively.

$$De = \sqrt{(x_i - s_i)^2 (y_i - t_i)^2} \tag{1}$$

$$v_{ij} = \frac{1}{N} \sum_{k=0}^{ni} X_{ij} \tag{2}$$

The Equations 3 and 4 shows formulae of DBSCAN's [11].

$$N_{EPS} = \left\{ \frac{qED}{dist(p,q)} < EPS \right\} \tag{3}$$

$$N_{EPS}(P) > MinPts \tag{4}$$

2.2. K-Nearest Neighbor Algorithm

K-Nearest algorithm, which is frequently used for classification problems. [12]. For a particular query point, distance metrics like Euclidean distances and Manhattan are essential in determining the closest points or groups. KNN is an instance-based learning technique that uses Euclidean distance to identify neighbors while keeping every sample used for classification. The majority vote among nearby samples is frequently used to determine the classification label. The fixed character of the variable K and biases resulting from non-uniform data distribution across categories present challenges that can affect model stability and cause overfitting. When the overall distance

traveled is more important, the Manhattan Distance Metric is used. A popular method for figuring out K is cross-validation, which isn't necessarily practical in circumstances like online categorization.

$$y(d_i) = arg_k^{max} \sum_{x_{jEKNN}} y(x_j, c_k) \quad (5)$$

$$y(d_i) = arg_k^{max} Sim\,(d_i, x_j)\, y(x_j, c_k) \quad (6)$$

$$Distance = \sum_{1}^{n} |p_i - q_i| \quad (7)$$

The formula is used to calculate accuracy, which is the percentage of right predictions.

$$Accuracy\ = \frac{TP + TN}{TP + FP + FN + TN} \quad (8)$$

For computation of Kappa statistics [13]

$$P(r) = (a + d)(a + b + c + d) \quad (9)$$

The value of Kappa can be defined as

$$K = Pr\ Pr\ (a)\ - \frac{Pr\ Pr\ (e)}{1} - Pr\ Pr\ (e) \quad (10)$$

For computation of sensitivity and specificity:

$$Sensitivity = \frac{TP}{TP + FN} \quad (11)$$

$$Specificity = \frac{TN}{TN + FP} \quad (12)$$

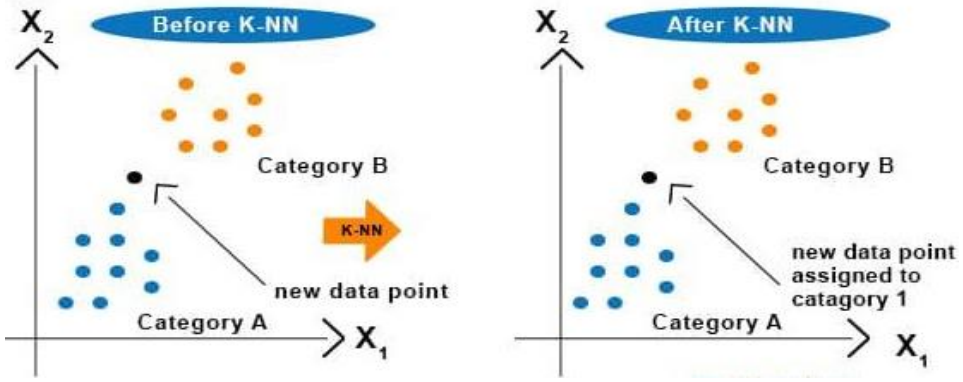

Figure 1. Visualization of the K-Nearest Neighbors (KNN) algorithm applied. (a) Before KNN: Original dataset with raw feature points and no classification boundaries. (b) After KNN: The dataset with classifications applied, showing decision boundaries.

2.3. Decision Tree with Random Forest

Recursive partitioning is the method by which Decision Trees (DT) create decision-making nodes and predicting leaves with the goal of minimizing standard variation by using a divide-and-conquer tactic. In order to find the best splits, one must search greedily until the termination requirements are satisfied [14]. The most important variable determines which node is the root, and the sum of the squared errors is minimized by the nodes that follow. Adopting Occam's razor, pruning and cross-validation are used for complexity evaluation, and smaller trees are favored to avoid overfitting. Ideal attributes for nodes can be found with the use of the Knowledge Acquisition Gini Index, which is a component of the attribute selection measure (ASM). Entropy is a crucial statistic that measures the unpredictability of data. The formula for calculating information gain is:

$$Information\ Gain = Entropy(S) - [(Weighted\ Avg) * Entropy(each\ feature)] \quad (13)$$

$$Entropy\ (S) = P(Yes).P(Yes)log2 - P(no)log2\ P(no) \quad (14)$$

$$Gini\ index = 1 - \sum_{j}^{2} p_j \quad (15)$$

The Gini index is used as a purity/impurity metric in the context of the CART algorithm [15]. It is a measure of statistical dispersion, quantifies the inequality of distribution within a dataset by calculating the degree of impurity or purity in classification tasks, with lower values indicating more homogenous groups. This approach generates binary splits, which advances the state of science, and the Random Forest ensemble method combines uncorrelated decision tree to improve accuracy [16,17].

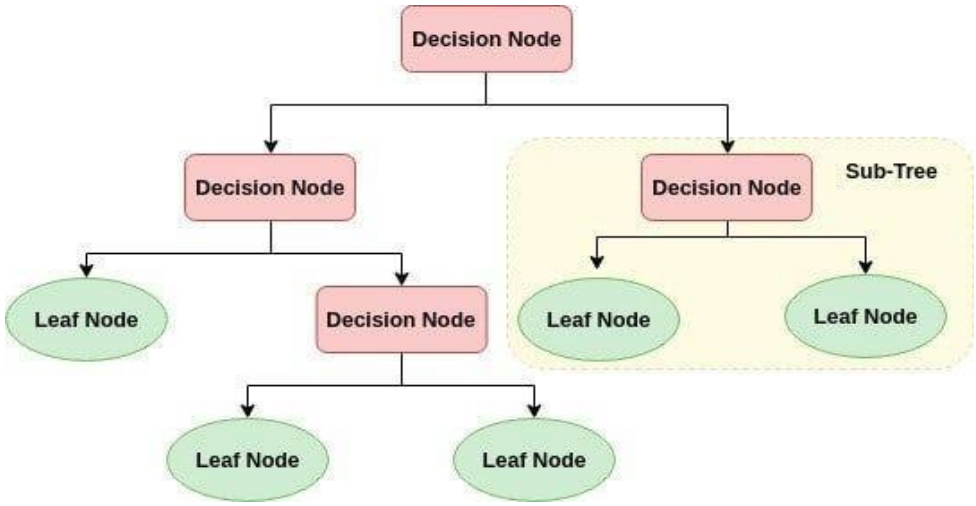

Figure 2. Decision Tree showing key components: Decision Node, where a decision is made based on input features; Leaf Node, which represents the outcome or classification result; and Sub-Tree, depicting a hierarchical structure of interconnected nodes and branches within the tree.

One prominent example of a flexible and intuitive machine learning method that can be used for problems involving regression as well as classification is the random forest classifier. Its strength is in using the bagging technique, which requires training many models, to create a combination of decision trees. The strengths of each individual tree are combined in this ensemble approach to improve the overall performance of the model. The versatility and ease of use of the random forest make it a useful tool in a variety of predictive modeling scenarios and highlights its ensemble technique also yields better results [18].

3. Results

3.1. Data Set and Pre-processing of Data

The dataset consists of 2111 instances with 17 features, taken from the open-source library UCI. The data set includes binary, binomial, and ordinal types. This data is well organized, with no missing values, and divided into five categories: basic information, family history, eatable habits, physical conditions, and harmful behavior. Figure 3 shows the categories of the data set.

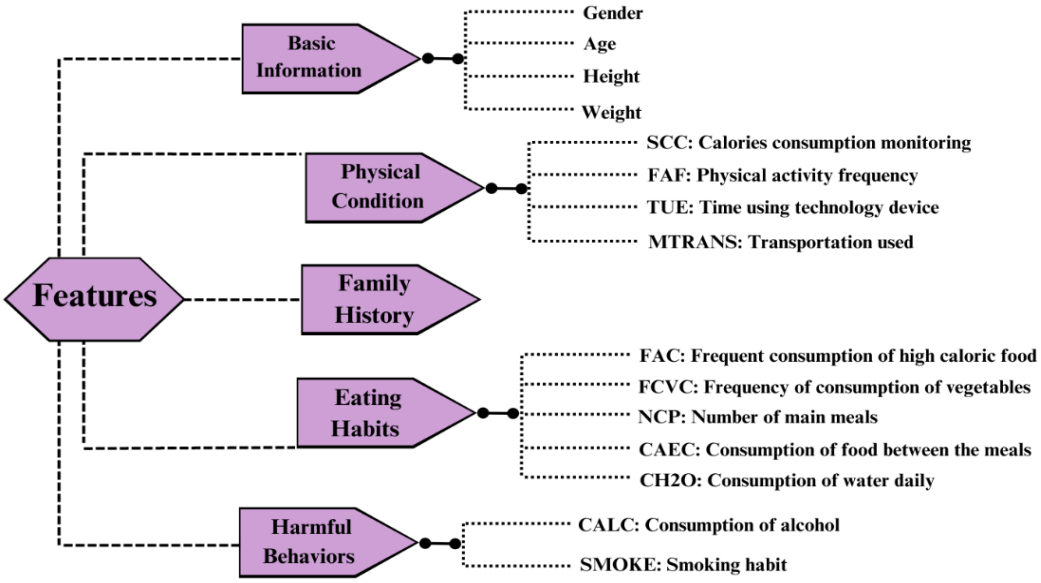

Figure 3. Overview of dataset divisions used in the analysis: (a) Basic Information (b) Family History (c) Eatable Habits (d) Physical Conditions (e) Harmful Behavior.

3.2. K-Mean Clustering

Unsupervised technique, K-mean clustering, is applied to the obesity data set with the labeled attribute, Use of Screen of Time. In this method, the value of k is 4, and the data set is clustered into 4 categories. In figure 4, the clusters are shown with different colors.

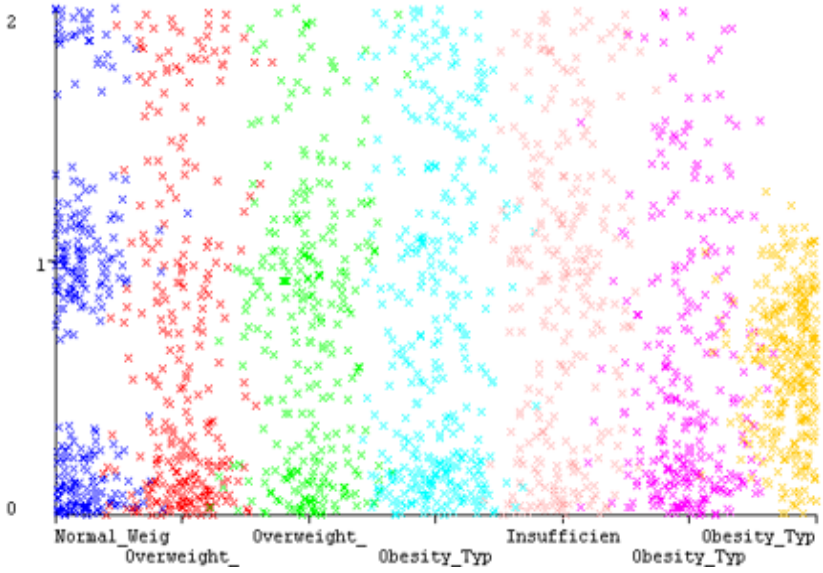

Figure 4. Dot Plot Matrix comparing the raw and synthetic Obesity datasets, illustrating feature distributions and relationships.

The initial points chosen for clusters, commonly referred to as centroids, play a crucial role in the outcome of the clustering process as shown in table 1. These initial centroids are either randomly selected from the dataset or determined through a more systematic approach such as the K-means initialization method. The selection of initial centroids significantly impacts the convergence rate and the final cluster quality.

Table 1. The Initial Points Chosen For Clusters 0, Clusters 1, Clusters 2, and Clusters 3.

Attribute	Gender	Age	Height	Weight	Family History	Fast Food Intake	Vegetable Consumption	Number of Meals Daily	Use of Food B/w Meals	Smoking	Use of Water	Calories Consumption	Physical Activity	Use of Technology	Alcohol Consumption	Type of Transport	Obesity Level
Cluster0	F	19	1.63	76	yes	No	3	3	Frq	yes	3	no	2	1	S	Auto	OvII
Cluster1	F	21.57	1.75	133.95	yes	yes	3	3	S	no	2.83	no	1.41	0.86	S	PT	ObIII
Cluster2	M	29	1.77	83	no	yes	1	4	Frq	no	3	no	0	1	no	Moto	OvI
Cluster3	M	28.36	1.79	112.73	yes	yes	1.99	3	S	no	2	no	0.29	1	S	PT	ObII

Notes: M = Male, F = Female, S = Sometimes, Frq = Frequently, PT = Public Transport, Ato = Automobile, Moto = Motor Bike, Ov = Overweight, and Ob = Obesity.

In Table 2, the comprehensive clusters are detailed along with their instances and associated values. Specifically, the values for technology use across different clusters are also mentioned. Notably, Cluster 2, which is predominantly male, demonstrates the highest value in screen time usage. This suggests that Cluster 2 has the most pronounced engagement with technology compared to other clusters, highlighting a potential association between gender and increased screen time within this dataset.

Table 2. Descriptive Statistics of Cluster 0, Cluster 1, Cluster 2, and Cluster 3 with Each Instances and their Values.

Attribute	Full Data	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Gender	Male	Female	Female	Male	Male
Age	24.3126	21.2279	25.5914	22.9032	25.5125
Height	1.7017	1.6338	1.6566	1.7042	1.7737
Weight	86.5861	52.3099	97.6555	78.0714	97.3083
Family History with Overweight	Yes	No	Yes	Yes	Yes
Fast Food Intake	Yes	Yes	Yes	Yes	Yes
Vegetable Consumption Frequency	2.419	2.4408	2.6855	2.278	2.2588
Number of Meals Daily	2.6856	2.7137	2.6751	2.6312	2.7275
Use of Food Between Meals	Sometimes	Frequently	Sometimes	Sometimes	Sometimes
Smoking	No	No	No	No	No
Use of Water	2.008	1.6477	2.0801	1.9337	2.1524
Calories Consumption Calculation	No	No	No	No	No
Physical Activity	1.0103	1.0639	0.7131	1.3637	0.9984
Use of Technology	0.6579	0.7559	0.5426	0.8644	0.5641
Alcohol Consumption	Sometimes	Sometimes	Sometimes	No	Sometimes
Type of Transport	PT	PT	PT	PT	PT
Obesity Level	Obesity I	Insufficient Weight	Obesity III	Normal Weight	Obesity II

Notes: PT = Public Transportation.

3.3. K-Nearest Neighbor Classifier

The several experiments for the K fold parameter are performed, from K = 2 to K = 10, and the best result is achieved at K = 9. The validity of the algorithm is checked by the K-fold cross validation method. The data set is divided into 70% for training and 30% for testing. The obtained accuracy of K nearest neighbor is 87.05%. The detail of predicted and actual classification is depicted in Table 3.

Table 3. The Confusion Matrix of KNN

	True False	True True	Class Precision
Predicted False	546	68	88.93%
Predicted True	14	5	26.32%
Class Recall	97.50%	6.85%	

3.4. Decision Tree with Random Forest

The random forest with cross-validation method gives correctly classified instances as 2019, and incorrectly classified instances as 92, which is 95.64% and 4.35%, respectively. In this classification, relative absolute error is 17.96% and root relative error squared error is 32.83%. Also, detailed accuracy by class is provided in Table 4. The confusion matrix of random forest is given in Table 5.

Table 4. Detail Accuracy of The Random Forest Method.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Detail Accuracy	0.944	0.027	0.844	0.944	0.891	0.875	0.993	0.952	NW
Detail Accuracy	0.900	0.010	0.932	0.900	0.916	0.903	0.995	0.975	Ov I
Detail Accuracy	0.952	0.008	0.952	0.952	0.952	0.944	0.997	0.987	Ov II
Detail Accuracy	0.960	0.003	0.985	0.960	0.973	0.967	0.999	0.997	Ob I
Detail Accuracy	0.949	0.001	0.996	0.949	0.972	0.968	0.999	0.995	IW
Detail Accuracy	0.987	0.001	0.993	0.987	0.990	0.988	1.000	0.999	Ob II
Detail Accuracy	0.997	0.001	0.997	0.997	0.997	0.996	1.000	1.000	Ob III
Weight Average	0.956	0.007	0.959	0.956	0.957	0.950	0.998	0.987	

Notes: NW = Normal Weight, Ov = Overweight Level, Ob = Obesity Type, IW = Insufficient Weight, and Av. = Weight Average.

A confusion matrix is a useful tool in machine learning for evaluating the performance of classification models by showing the true and false positives and negatives. In the context of body weight classification, the Body Mass Index (BMI) ranges are defined as follows: Normal weight falls within a BMI of 18.5 to 24.9, while those classified as Overweight Level I have a BMI ranging from 25.0 to 29.9. Overweight Level II includes individuals with a BMI between 30.0 and 34.9. Obesity is further categorized into three types: Type I, which covers a BMI from 35.0 to 39.9; Type II, with a BMI ranging from 40.0 to 44.9; and Type III, for a BMI of 45.0 or higher. Conversely, insufficient weight is indicated by a BMI less than 18.5.

Table 5. Confusion Matrix of The Random Foresting.

a	b	c	d	e	f	g	Classified as
271	13	2	0	1	0	0	a = Normal Weight
26	261	3	0	0	0	0	b = Overweight Level I
6	5	276	3	0	0	0	c = Overweight Level II
3	1	9	337	0	1	0	d = Obesity Type I
14	0	0	0	258	0	0	e = Insufficient weight
1	0	0	2	0	293	1	f = Obesity Type II
0	0	0	0	0	1	323	g = Obesity Type III

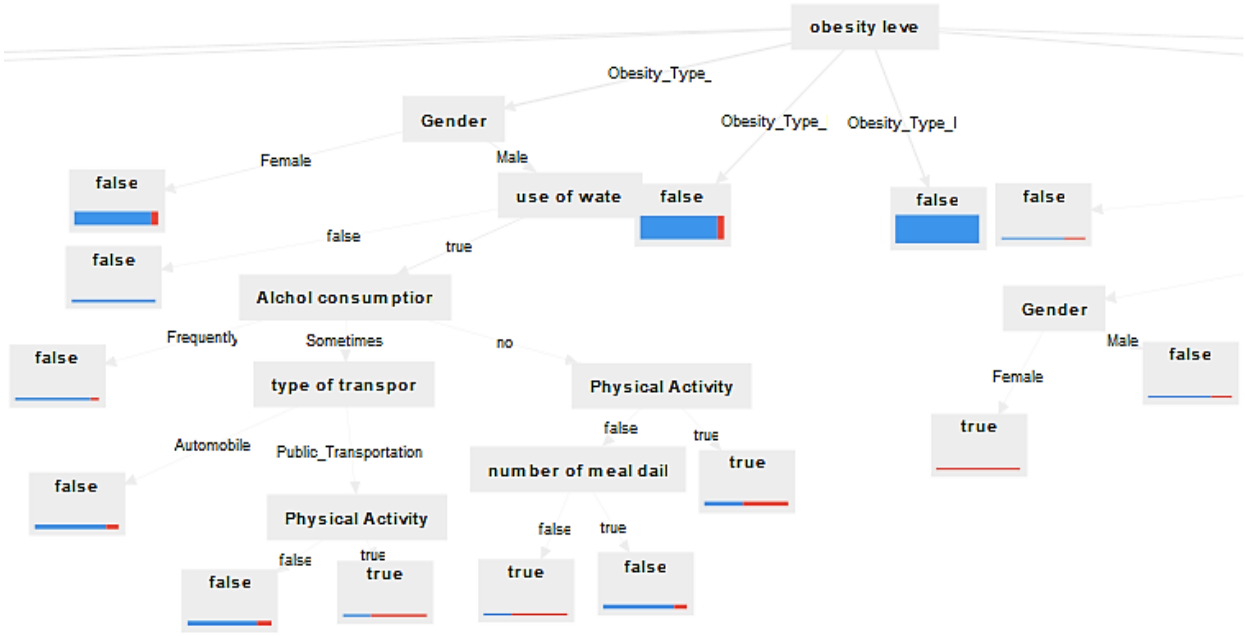

Figure 5. A Segment of The Decision Tree, With Obesity Levels as The Top Attribute and Physical Activity at The Bottom.

In figure 5, the small piece of the decision tree is depicted, starting from attribute obesity levels and ending on physical activity. Each branch of decision tree is repressing the different levels of obesity then going down on basis of importance and strong correlation of other features with obesity.

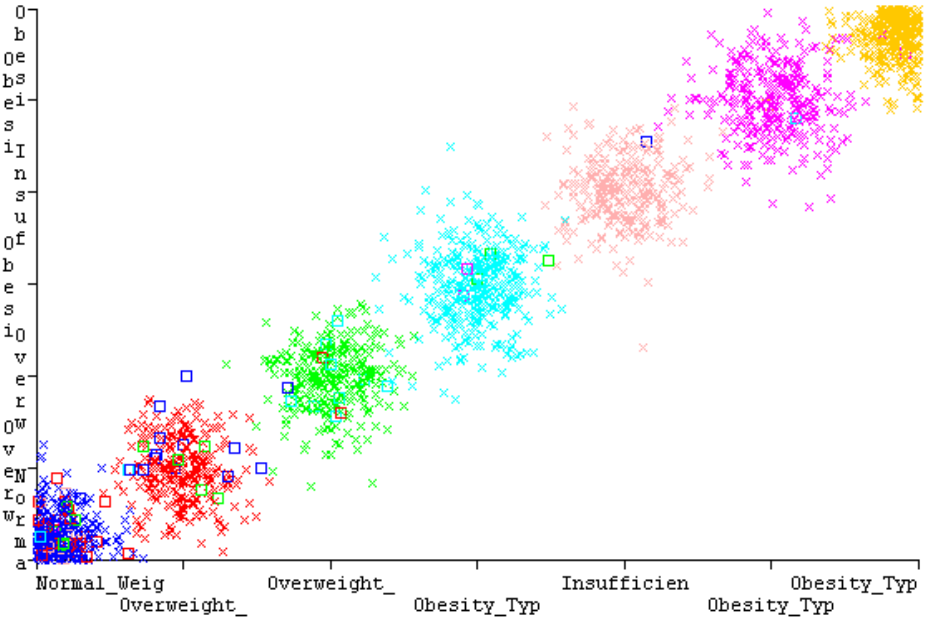

Figure 6. Classification of Obesity Types With Random Foresting Where Orange Indicates Insufficient Weight, Pink Indicates Type II Obesity, Yellow Indicates Type III Obesity, Blue Indicates Normal Weight, Red Indicates Level I and II Overweight, And Light Blue Indicates Type I Obesity.

The results of the decision tree with random forest are shown in figure 6. It can be seen that normal weight and obesity types are inter-relatable with each other and gradually increasing. It is also showing that there are some instances with different colors that are specified incorrectly.

5. Conclusions

Obesity is a major global health concern, affecting over 20 billion people worldwide and its prevalence increasing at alarming rates. This is a complex issue affected by various factors, like social dynamics, genetic behavior, and environmental factors. Along with these factors, a sedentary lifestyle is also the most significant contributor, which is due to the lack of physical activity. The period of inactivity is led by factors like use of mobile phones, laptops, and prolonged screen time. Machine learning techniques are contributing the main role to tackle issues in health, predict future outcomes, and provide insightful clusters. To address the global epidemic of obesity requires multi-faceted approaches. The machine learning approaches, K-mean clustering, K-NN classifier, and Decision Tree with Random Foresting have provided very high accuracy rates up to 95.64%. These techniques have established a very strong relationship between use of screen time and obesity.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lissak G. Commentary on the article: "Adverse physiological and psychological effects of screen time on children and adolescents -- Literature review and case study" - Environmental Research 164 (2018) 149-157. J Ment Health Clin Psychol. 2018;2(3). doi:10.29245/2578-2959/2018/3.1134
2. Mukhtar S, Zubair R, Masroor S, Jaffar S, Anjum A, Naqvi S. Effect Of Screen-Time on Obesity: A Cross-Sectional Study. Pakistan Journal of Public Health. 2023;13(2). doi:10.32413/pjph.v13i2.1114
3. Stiglic N, Viner RM. Effects of screentime on the health and well-being of children and adolescents: A systematic review of reviews. BMJ Open. 2019;9(1). doi:10.1136/bmjopen-2018-023191
4. Tambalis KD, Panagiotakos DB, Psarra G, Sidossis LS. Insufficient sleep duration is associated with dietary habits, screen time, and obesity in children. Journal of Clinical Sleep Medicine. 2018;14(10). doi:10.5664/jcsm.7374
5. Marttinen R, Vernikoff L, Phillips S, Fletcher N. Physical Activity, Screen Time, and Obesity. Calif J Health Promot. 2017;15(1). doi:10.32398/cjhp.v15i1.1886
6. Robinson TN, Banda JA, Hale L, et al. Screen media exposure and obesity in children and adolescents. Pediatrics. 2017;140. doi:10.1542/peds.2016-1758k
7. Shamshirband S, Fathi M, Dehzangi A, Chronopoulos AT, Alinejad-Rokny H. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. J Biomed Inform. 2021;113. doi:10.1016/j.jbi.2020.103627
8. Alkhalaf M, Yu P, Shen J, Deng C. A review of the application of machine learning in adult obesity studies. Applied Computing and Intelligence. 2022;2(1). doi:10.3934/aci.2022002
9. Sinaga KP, Yang MS. Unsupervised K-means clustering algorithm. IEEE Access. 2020;8. doi:10.1109/access.2020.2988796
10. Schott M. K-Means Clustering Algorithm for Machine Learning . Capital One Tech on Medium.
11. Chen Y, Zhou L, Bouguila N, Wang C, Chen Y, Du J. BLOCK-DBSCAN: Fast clustering for large scale data. Pattern Recognit. 2021;109. doi:10.1016/j.patcog.2020.107624
12. Sharma M. Generalized K-Nearest Neighbour Algorithm- A Predicting Tool. International Journal of Advanced Research in Computer Science and Software Engineering. 2013;3(11).
13. McHugh ML. Interrater reliability: The kappa statistic. Biochem Med (Zagreb). 2012;22(3). doi:10.11613/bm.2012.031
14. Quinlan JR. Induction of Decision Trees. Mach Learn. 1986;1(1). doi:10.1023/a:1022643204877
15. Barot PA, Jethva HB. Mgini-improved decision tree using minority class sensitive splitting criterion for imbalanced data of covid-19. Journal of Information Science and Engineering. 2021;37(5). doi:10.6688/jise.202109_37(5).0008
16. Talekar B. A Detailed Review on Decision Tree and Random Forest. Biosci Biotechnol Res Commun. 2020;13(14). doi:10.21786/bbrc/13.14/57
17. Fratello M, Tagliaferri R. Decision trees and random forests. In: Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. Vol 1-3. ; 2018. doi:10.1016/b978-0-12-809633-8.20337-3
18. Wu DJ, Feng T, Naehrig M, Lauter K. Privately Evaluating Decision Trees and Random Forests. Proceedings on Privacy Enhancing Technologies. 2016;2016(4). doi:10.1515/popets-2016-0043
19. LeCroy, M.N., Kim, R.S., Hanna, D.B., Perreira, K.M., Gallo, L.C., Llabre, M.M., Van Horn, L., Daviglus, M.L., Talavera, G.A., Sotres-Alvarez, D., & Isasi, C.R. (2023). Abstract P128: Predicting BMI Percentile in Hispanic/Latino Youth Using a Machine Learning Approach: Findings From the Hispanic Community Children's Health Study/Study of Latino Youth. Circulation.
20. Hammond R, Athanasiadou R, Curado S, Aphinyanaphongs Y, Abrams C, Messito MJ, Gross R, Katzow M, Jay M, Razavian N, Elbel B. Predicting childhood obesity using electronic health records and publicly available data. PLoS One. 2019 Apr 22;14(4):e0215571. doi: 10.1371/journal.pone.0215571. Erratum in: PLoS One. 2019 Oct 7;14(10):e0223796. doi: 10.1371/journal.pone.0223796. PMID: 31009509; PMCID: PMC6476510.
21. Lotfata, A., Georganos, S., Kalogirou, S., & Helbich, M. (2022). Ecological Associations between Obesity Prevalence and Neighborhood Determinants Using Spatial Machine Learning in Chicago, Illinois, USA. ISPRS Int. J. Geo Inf., 11, 550.