Obesity Level Prediction based on Machine Learning Techniques

1st Mingyu Wan
Department of Chemical Engineering
University of Massachusetts - Lowell
MA, USA
Mingyu_Wan@student.uml.edu

2nd Eric Zabele
Department of Computer Science
University of Massachusetts - Lowell
MA, USA
eric_zabele@student.uml.edu

3rd Amin Majdi
Department of Computer Science
University of Massachusetts - Lowell
MA, USA
Amin_Majdi@student.uml.edu

***Abstract*—Obesity has been on the rise for the past few decades and is now a challenge for all populations worldwide, regardless of age or gender. Obesity causes many cardiovascular diseases, so an efficient and accurate method to predict obesity by the basic state of the body and lifestyle habits is necessary. To explore the role of machine learning techniques in obesity prediction, we compared six different algorithms, i.e., K-Nearest Neighbors, Naive Baye, Decision Trees, random forests, Support Vector Machines and Artificial Neural Network. The results show that, with Likert scale data, decision trees with increasing depth, artificial neural networks, and support vector machines using a linear kernel and a regularization parameter of 2 performed the best.**

***Index Terms*—Machine Learning, Data mining, Support Vector Machine, Decision Tree, Obesity, Prediction, Naive Bayes, Artificial Neural Networks, Random Forest, K Nearest Neighbors.**

I. Introduction

With the advances in technology and agriculture in the past decades, extra food is available more than in history. As a result, the obesity epidemic became a universal problem [1]. Obesity is one of the primary causes of many different illnesses like diabetes and cancers [2]. Obesity refers to the amount of extra body fat accumulated around the belly and is very harmful to overall health [3]. There are many different risk factors like genetics, eating habits, social habits, etc., that can increase the chance of obesity in adults. According to the Global Burden of Disease study, obesity was the leading cause of 4.7 million death in 2017 [4] which shows the importance of this global issue.

Despite all the efforts to control obesity globally [5], there is no guaranteed solution for this universal issue. None of the available methods successfully predicted obesity in its early stage [6]. As a result, Obesity Level prediction became one of the popular and challenging topics during the past decades due to its importance. Thus, the number of publications on this topic was significant. Researchers mainly focused on many different factors that can affect the obesity level in the adult population. Most of these researchers used machine learning algorithms to analyze various databases and investigate factors that can affect obesity levels [7] [8].

In this study, the main objective is to investigate the performance of five different machine learning algorithms on an obesity database and compare their result and build a optimized model to predict obesity level of adults based on their features. This paper is structured in the following manner. Section 2 elaborates on the related works. Section 3 describes proposed techniques and data set structure. Section 4 focuses on results and discussion about the accuracy of the findings, and Section 5 expounds on discussion highlights.

II. Related Works

A. Obesity Prediction Using Machine Learning

Previous research has shown that machine learning algorithms can successfully make models for biomedical applications[13]. Obesity prediction is one area in which machine learning algorithms have been used frequently in the past decade[14]. To conduct a machine learning classification, the researchers tended to find various features affecting obesity, and each research investigated the influence of some of the selected features, from behavioral features like smoking habits [15] to biological and social features [16] [17].

Alongside the feature selection, choosing a well Machine learning algorithm is also controversial. Researchers used various single and hybrid ML methods to predict obesity levels in different age groups. According to the review done by Mahmood Safaei et al.[7], The artificial neural network was the most popular machine learning algorithm among researchers. Meanwhile, Support vector machines [18] [19] [20] and decision trees [21] [22] [23] were also used to predict obesity levels frequently.

B. Results On The Current Database

This research uses the data set gathered from countries like Mexico, Peru, and Colombia in the 14- 61year age group with various physical conditions and different eating habits. Some other works used the mentioned dataset. Yaren Celic et al. [9] used the dataset to investigate the effects of the used features on classification success. There are 16 dependent features in the dataset. They used the 12 most important features to reveal that the success rate can be high, even without less essential features. The success rate for 13 different algorithms was examined and was compared Using these 12 features(6 decision tree algorithms, 6 SVM algorithms, and the artificial Neural network algorithm). Cubic SVM had the best

TABLE I: The results for the previous works on the dataset

Study	Method	Number of features	Accuracy (%)
Yaren Celic et al. [9]	Fine Tree	12	93.1
	Medium Tree	12	81.9
	Coarse Tree	12	60.4
	Linear SVM	12	96.0
	Quadratic SVM	12	97.2
	Cubic SVM	12	**97.8**
	Fine Gaussian SVM	12	85.6
	Medium Gaussian SVM	12	95.4
	Coarse Gaussian SVM	12	88.4
	Boosted Tree	12	90.4
	Bagged Tree	12	95.4
	RUSBoosted Tree	12	83.2
	Artificial Neural Network	12	96.5
Asma Alqahtani et al. [10]	Random Forest	16	96.70
	Multi-Layer Perception	16	95.06
Satvik Garg et al. [11]	Random Forest	16	86
	Decision Tree	16	76
	Extra Tree	16	85
	KNN	16	82
	XGBoost	16	85

performance among the algorithms used by its 97.8% accuracy rate. They conclude that faster results can be obtained by fewer features. This elimination can decrease calculation costs and save the data collection time and effort.

Asma Alqahtani et al. [10] applied two machine learning techniques, named random forest and Multi-Layer Perception(MLP), to the Database to classify these data and build a model to predict obesity levels. They concluded that random forest had a better performance than MLP at the early stage with an accuracy of 96.70%. In contrast to Celic research, they claimed that using these two algorithms, the best performance was established when they used all 16 dependent features in their calculations.

Satvik Garg et al. [11] also used the database to provide a framework that uses machine learning algorithms, namely, Random Forest, Decision Tree, XGBoost, Extra Trees, and KNN to train three models to predict obesity levels, Bodyweight, and fat percentage levels. In this research, the random forest model using grid search and Tpot classifier outperformed the other methods with 86% accuracy rate. Furthermore, They made a website for obesity prediction using these models to offer diet plans.

In brief, you can find the results of the three studies in **Table 1**.

III. Methodology

A. Data Set

The dataset was created by the information collection from a survey. These questions included [24]: What is your gender? What is your age? What is your height? What is your weight? Has a family member suffered or suffers from overweight? Do you usually eat high caloric food frequently? Do you usually eat vegetables in your meals? How many main meals do you have daily? Do you eat any food between meals? Do you smoke? How much water do you drink daily? Do you monitor the calories you eat daily? How often do you have physical activity? How much time do you use technological devises such as cell phone, videogames, television, computer and others? How often do you drink alcohol? Which transportation do you usually use? These 16 questions could serve as the attributes in the dataset. All data was labeled the obesity level according to body mass index (BMI) formula from WTO:

$$BMI = \frac{Weight}{Height^2}$$

where weight and height could be obtained from that survery. WTO also gives a BMI classification of obesity level shown in the **Table 2**. All the datapoints are from young undergraduate students between 18 and 25 years old from Colombia, Mexico and Peru. The size of this sample was 712 including 423 men and 388 women.

TABLE II: BMI classification

BMI Classification	
Normal	Less than 18.5
Overweight	18.5 to 24.9
ObesityI	25.0 to 29.9
ObesityII	30.0 to 34.9
ObesityIII	35.0 to 39.9

B. Algorithms

1) K Nearest Neighbors: The k-nearest neighbor (KNN) algorithm is a supervised learning that can be used for classification and regression. Given a training dataset, for new input datapoint, find the k points in the training dataset that are nearest to the new data. That new datapoint will be classified to the class that has the majority of the k points (**Fig. 1**). KNN is usually used in character recognition, text classification, image recognition, etc.

2) Naive Bayes: As a classification algorithm with high accuracy and speed, the core theory of Naive Bayes is the following Bayesian theory

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Bayesian theory describes the probability of the occurrence of event A, given that B is known. The prerequisite of Naive Bayes is that the attributes are independent of each other. The precision of classification becomes higher as the dataset satisfies the independence assumption better.

3) Decision Tree Classifier: A decision tree is considered as a classification procedure including a root node, a couple of child node and leaf node (**Fig. 2**); child nodes denote the tests in the attributes and leaf notes represent the result of the decisions. Decision Tree could be used in many areas such as medical diagnoses, radar signal and text recognition with high performance [25]. Decision tree can be performed by various algorithms such as ID3, C4.5 and RandomForest etc.

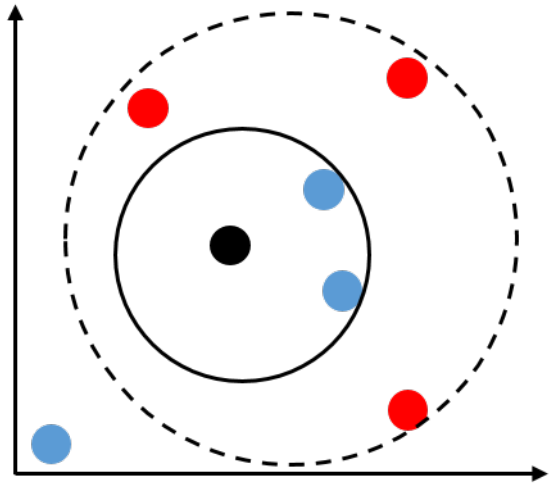


Fig. 1: SVM scheme

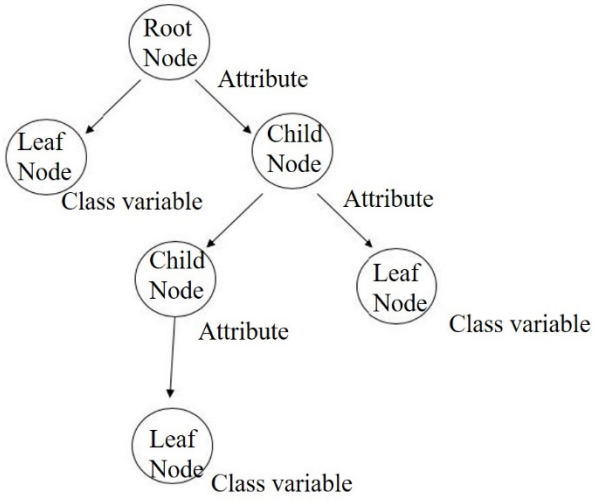


Fig. 2: Decision Tree Structure

4) Support Vector Machines: Support Vector Machines (SVM) is a supervised model used for classification (**Fig. 3**). An SVM uses support vectors to define a decision boundary. Classifications are made by comparing unlabeled points to that decision boundary. SVM has been applied in many fields such as text classification [26] and human activity recognition [27]. The major advantage of SVMs is the available of powerful tools algorithms to solve the problem efficiently and quickly [28].

5) Artificial Neural Networks: Artificial Neural Networks (ANN) is an operational model consisting of a large number of nodes (or neurons) which are connected by links. Each link as an associated weight and an activation level. Each node has an input function, an activation function and an output (**Fig. 4**).

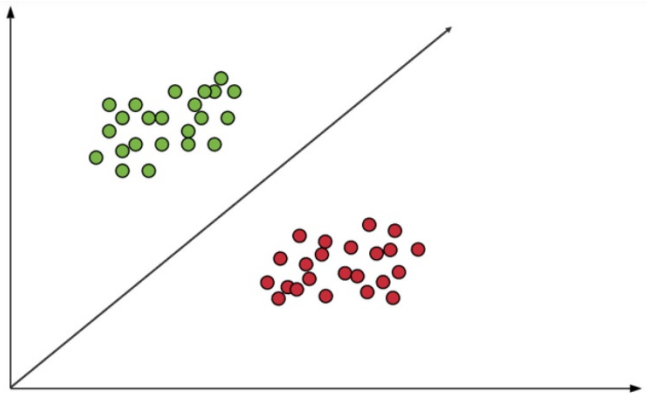


Fig. 3: SVM scheme

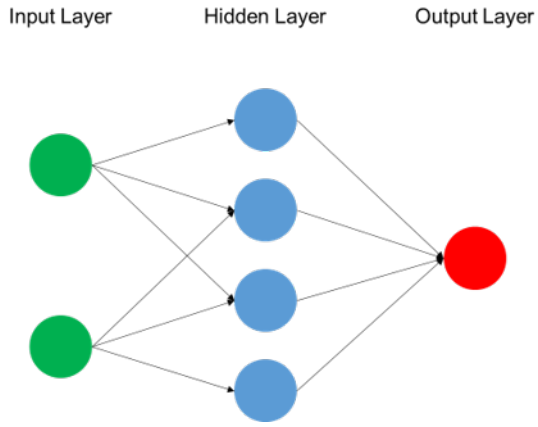
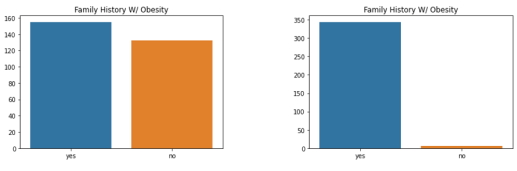


Fig. 4: ANN scheme

IV. Results and Discussion

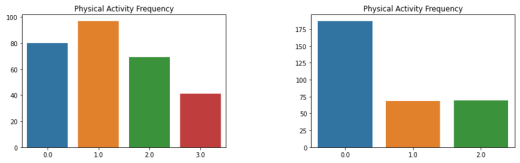
A. Exploratory Data Analysis

There are sixteen features: gender, age, height, weight, family history of obesity, consumption of high caloric foods frequency, vegetable consumption frequency, number of meals daily, consume food between meals, smoke, water consumption, monitoring of calorie consumption, physical activity frequency, time using technology, alcohol, and mode of transportation. The surveyors believed these to be the most contributing factors into one's obesity level. With the exception of age, height, and weight, all features were categorical in a binary or likert scale nature. There were 2111 entries which about 75% of was synthesized.



(a) Normal Weight (b) Obesity Level I

Fig. 5: Family history of obesity comparison



(a) Normal Weight (b) Obesity Level III

Fig. 6: Physical activity comparison

Figures 5 and 6 display the differences between a normal, healthy, weighted group against a level of obesity.

Upon looking at the trends, as obesity level increases, it is clear to see that family history of obesity was a strong indicator of one's own obesity level. There were many subtle differences in the trends between groups, as obesity level rose, the food intake would also increase. From intuition, a sedentary, high-caloric lifestyle would suggest health level would indicate obesity and we are hoping our chosen algorithms would indicate the same thing. In this research, KNN, Decision Trees, Multi-Layer perceptron, Random Forest, Naive Bayes, and Support vector machines were trained evaluated on their accuracy to label individuals appropriately to their category of obesity. Seven levels were outlined by the researchers and are referred to as normal, insufficient weight, overweight I and II, and obesity I, II, and III.

B. Algorithm Performances

In this subsection we will briefly discuss each algorithms performance. Before training and testing our data, we first had to perform some preprocessing to allow our models to function as we desired. The numerical data was not in the form of likert scale even though it was supposed to be. To overcome this issue, we rounded the values to their nearest whole number. We normalized age, height, and weight to avoid them providing too much influence unto our models. The categorical data in the dataset was factorized as our functions required it. After doing this we split our data 80-20 evenly to train and evaluate our models. We also tested these algorithms while dropping the height, weight, and age features. The results discussed below are using all 16 features. The first algorithm used to train models on this dataset was K-nearest neighbors. Over the course of performing this research, it was determined that we should explore the performance knn had with various different k values. Figure 7 shows how with an increasing number of k, the performance of the algorithm began to decline.

From this research it was consistently shown that a k-value equal to 1 would result with the highest accuracy with a score of about 79.0%. This result was repeatable over many different runs with different train - test samples and would not fluctuate too far from 80%. We speculate that with a low k value, individuals who have very similar lifestyles have similar obesity levels and as such, would lead to a more correct prediction.

The use of the decision tree was also explored on this dataset along with how depth impacted performance. As figure

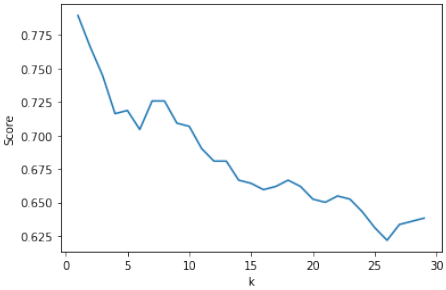
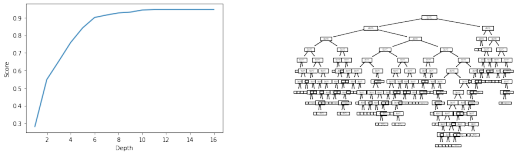


Fig. 7: k-value performance



(a) D-Tree max depth performance comparison (b) Resulting decision tree with max depth 9

Fig. 8: Decision Tree Results

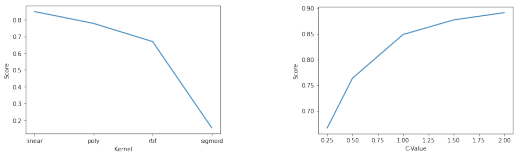
8 shows, with increasing max depth, performance increased until about a max depth of 11 where performance began to drop. Performance of decision trees began to even out around a max depth usage of 7.

We speculate that because of the wide range of surveys, certain features would be much more informative and thus increase the algorithms ability to predict obesity. In the same respect, the algorithm needed just enough categories to really predict well. With a max depth parameter of 11, our decision trees was accurate about 94.6% of the time.

In the beginning of our research, our SVM models were performing poorly so it was decided to just explore a large range of parameters. As it turns out, linear kernel performed the best with this dataset and would explain our prior results because of our original belief that and RBF kernel would be superior as we thought it could deal with higher dimensions better. However, our testing would consistently show us that linear outperformed RBF.

We subsequently tested a small range of regularization parameter values, known as the "C" parameter in the scikit library. We believe that the SVM performed better with a higher regularization parameter because L2-regularization imposes a bigger penalty to those points that violate the margins.

We also tested ANN, Random Forest, and naive bayes. Our



(a) SVM kernel performance (b) Linear Kernel C-Value comparison

Fig. 9: Physical activity comparison

multi-layer perceptron model performed great achieving an accuracy consistently in the 90% range with our best being at 92.7%. We trained our network using lbfgs optimization, RELU activation function, a learning rate of 0.001, 3 hidden layers consisting of 16, 11, and 6 neurons. We did no parameter tuning for naive bayes or random forest however we achieved accuracies of 61.5% and 95.5% respectively.

C. Discussion

In comparison with other research, our performance fared quite well. Table III shows a direct comparison of ours with aforementioned research on this dataset. Compared with the best results from this table, our Random Forest and artificial neural network models performed about as well as Alqahtani et al. Our SVM performed slightly worse than Celic et al. while our KNN achieved similar results to that of Garg et al. Our decision trees excelled, achieving high results without utilizing an ensemble method.

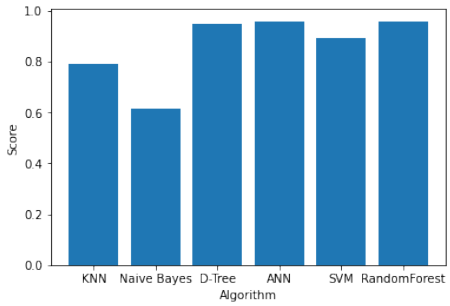


Fig. 10: Comparing algorithm performance with Age, Weight, Height

However we do note, the difference in the number of features used and we shall discuss further. In comparison with ourselves, it was found that if we left age, height, and weight out of our dataset, limiting our models to the other 13 features, our models would all perform in the 70% range meaning a 20% dropoff from our top performers. This could indicate that our models, when trained with that information, was leaning heavily on what we might consider highly correlated data. We suspect it is most likely weight as it could be argued that it is highly correlated with obesity and is why our ANN performed so poorly without it yet great with it as it most likely applied a high weight to that feature. Overall our performance as researchers improved over the course of this research, increasing our model accuracies to achieve good results towards the end.

V. Conclusion

In conclusion, obesity has impacted the lives of many individuals around the world and unfortunately leads to other life and health complications as well. The aim of this project was to identify what classification algorithms perform well using the chosen dataset. The results indicate that for using Likert scale data, decision trees with increasing depth, artificial neural networks, and support vector machines using a linear kernel and a regularization parameter of 2 performed the best. For future research, feature importance should be done to see if the machine learning algorithms can identify what features are mostly closely linked with obesity and compare those results to what we know from biology.

TABLE III: Similar previous work vs us

Study	Method	Number of features	Accuracy (%)
Yaren Celic et al. [9]	Fine Tree	12	93.1
	Medium Tree	12	81.9
	Coarse Tree	12	60.4
	Linear SVM	12	96.0
	Boosted Tree	12	90.4
	Bagged Tree	12	95.4
	RUSBoosted Tree	12	83.2
	Artificial Neural Network	12	96.5
Asma Alqahtani et al. [10]	Random Forest	16	96.70
	Multi-Layer Perception	16	95.06
Satvik Garg et al. [11]	Random Forest	16	86
	Decision Tree	16	76
	Extra Tree	16	85
	KNN	16	82
Us	KNN	16	79.0
	Naive Bayes	16	61.5
	Decision Tree	16	94.6
	Multi-Layer Perceptron	16	92.7
	Linear SVM	16	89.1
	Random Forest Default	16	95.5
	KNN	13	67.6
	Naive Bayes	13	52.5
	Decision Tree	13	72.3
	Multi-Layer Perceptron	13	65.5
	Linear SVM	13	59.6
	Random Forest Default	13	74.7

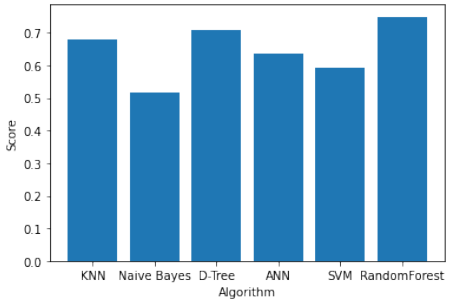


Fig. 11: Comparing algorithm performance w/o age, weight, height

References

[1] OMS. Organización mundial de la Salud. 2016. http://www.who.int/mediacent re/factsheets/fs311/es/. 1955.

[2] H.B.Hubert, M.Feinleib,P.M.McNamara, and W. P. Castelli, "Obesity as an independent risk factor for cardiovascular disease: A 26-year follow-up of participants in the Framingham Heart Study," Circulation, vol. 67, no. 5, pp. 968–977, 1983, doi: 10.1161/01.CIR.67.5.968.

[3] WHO, Obesity and overweight, 2020.[Enlínea]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[4] Hannah Ritchie and Max Roser (2017) - "Obesity". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/obesity' [Online Resource]

[5] O'Meara, S., Riemsma, R., Shirran, L., Mather, L. and Ter Riet, G., 2002. The clinical effectiveness and cost-effectiveness of sibutramine in the management of obesity: a technology assessment. Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews [Internet].

[6] Ivanescu, A.E., Li, P., George, B., Brown, A.W., Keith, S.W., Raju, D. and Allison, D.B., 2016. The importance of prediction model validation and assessment in obesity and nutrition research. International journal of obesity, 40(6), pp.887-894.

[7] Safaei, M., Sundararajan, E.A., Driss, M., Boulila, W. and Shapi'i, A., 2021. A systematic literature review on obesity: Understanding the causes consequences of obesity and reviewing various machine learning approaches used to predict obesity. Computers in biology and medicine, p.104754.

[8] DeGregory, K.W., Kuiper, P., DeSilvio, T., Pleuss, J.D., Miller, R., Roginski, J.W., Fisher, C.B., Harness, D., Viswanath, S., Heymsfield, S.B. and Dungan, I., 2018. A review of machine learning in obesity. Obesity reviews, 19(5), pp.668-685.

[9] Celik, Y., Guney, S. and Dengiz, B., 2021, July. Obesity Level Estimation based on Machine Learning Methods and Artificial Neural Networks. In 2021 44th International Conference on Telecommunications and Signal Processing (TSP) (pp. 329-332). IEEE.

[10] Alqahtani, A., Albuainin, F., Alrayes, R., Alyahyan, E. and Aldahasi, E., 2021. Obesity Level Prediction Based on Data Mining Techniques. International Journal of Computer Science Network Security, 21(3), pp.103-111.

[11] Garg, S. and Pundir, P., 2021, August. MOFit: A Framework to reduce Obesity using Machine learning and IoT. In 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO) (pp. 1733-1740). IEEE.

[12] Cui, T., Chen, Y., Wang, J., Deng, H. and Huang, Y., 2021, May. Estimation of Obesity Levels Based on Decision Trees. In 2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM) (pp. 160-165). IEEE.

[13] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H., 2018. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, p.515.

[14] Acharjee, A., Ament, Z., West, J.A., Stanley, E. and Griffin, J.L., 2016. Integration of metabolomics, lipidomics and clinical data using a machine learning method. BMC bioinformatics, 17(15), pp.37-49.

[15] Cheng, H., Montgomery, S., Green, A. and Furnham, A., 2020. Biomedical, psychological, environmental and behavioural factors associated with adult obesity in a nationally representative sample. Journal of Public Health, 42(3), pp.570-578.

[16] Kadouh, H.C. and Acosta, A., 2017. Current paradigms in the etiology of obesity. Techniques in Gastrointestinal Endoscopy, 19(1), pp.2-11.

[17] Sartorius, B., Veerman, L.J., Manyema, M., Chola, L. and Hofman, K., 2015. Determinants of obesity and associated population attributability, South Africa: Empirical evidence from a national panel survey, 2008-2012. PloS one, 10(6), p.e0130218.

[18] Montañez, C.A.C., Fergus, P., Hussain, A., Al-Jumeily, D., Abdulaimma, B., Hind, J. and Radi, N., 2017, May. Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 2743-2750). IEEE.

[19] Dunstan, J., Aguirre, M., Bastías, M., Nau, C., Glass, T.A. and Tobar, F., 2020. Predicting nationwide obesity from food sales using machine learning. Health informatics journal, 26(1), pp.652-663.

[20] Wang, H.Y., Chang, S.C., Lin, W.Y., Chen, C.H., Chiang, S.H., Huang, K.Y., Chu, B.Y., Lu, J.J. and Lee, T.Y., 2018. Machine learning-based method for obesity risk evaluation using single-nucleotide polymorphisms derived from next-generation sequencing. Journal of Computational Biology, 25(12), pp.1347-1360.

[21] Fernández-Navarro, T., Díaz, I., Gutiérrez-Díaz, I., Rodríguez-Carrio, J., Suárez, A., Clara, G., Gueimonde, M., Salazar, N. and González, S., 2019. Exploring the interactions between serum free fatty acids and fecal microbiota in obesity through a machine learning algorithm. Food Research International, 121, pp.533-541.

[22] Taghiyev, A., Altun, A.A. and Caglar, S., 2020. A Hybrid Approach Based on Machine Learning to Identify the Causes of Obesity. Journal of Control Engineering and Applied Informatics, 22(2), pp.56-66.

[23] de Moura Carvalho, L., Furtado, V., de Vasconcelos Filho, J.E. and Lamboglia, C.M.G.F., 2016, September. Using Machine Learning for Evaluating the Quality of Exercises in a Mobile Exergame for Tackling Obesity in Children. In Proceedings of SAI Intelligent Systems Conference (pp. 373-390). Springer, Cham.

[24] E. De-La-Hoz-Correa, F. E. Mendoza-Palechor, A. De-La-Hoz-Manotas, R. C. Morales-Ortega, and S. H. Beatriz Adriana, "Obesity Level Estimation Software based on Decision Trees," Journal of Computer Science, vol. 15, no. 1, pp. 67-77, 2019.

[25] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol. 21, no. 3, pp. 660-674, 1991.

[26] T. Joachims, "Text categorization with support vector machines : Learning with many relevant features," Lecture notes in computer science, pp. 137-142, May 1998.

[27] K. Youngwook and L. Hao, "Human Activity Classification Based on Micro-Doppler Signatures Using a Support Vector Machine," IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 5, pp. 1328-1337, 2009.

[28] E. de la Hoz, E. de la Hoz, A. Ortiz, J. Ortega, and A. Martínez-Álvarez, "Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps," Knowledge-Based Systems, vol. 71, pp. 322-338, 2014.