Machine Learning Based Adult Obesity Prediction

K.Nirmala Devi
Associate Professor: CSE. Kongu Engineering College
Erode, India
k_nirmal.cse@kongu.edu

N.Krishnamoorthy
Associate Professor: CSE. Kongu Engineering College
Erode, India
nmoorthy@kongu.ac.in

P.Jayanthi
Associate Professor: CSE. Kongu Engineering College
Erode, India
jayanthime@kongu.ac.in

S.Karthi
UG Scholar: CSE. Kongu Engineering College
Erode, India
karthis.18cse@kongu.edu

T.Karthik
UG Scholar: CSE. Kongu Engineering College
Erode, India
karthikt.18cse@kongu.edu

K.Kiranbharath
UG scholar: CSE. Kongu Engineering College
Erode, India
kiranbharathk.18cse@kongu.edu

2022 International Conference on Computer Communication and Informatics (ICCCI) | 978-1-6654-8035-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCCI54379.2022.9740995

Abstract— **Obesity is a worldwide health problem that affects people of all ages and genders. Obese people have a higher risk of several major diseases and health disorders, such as hypertension, dyslipidemia, coronary heart disease, stroke, and so on. The proposed technique combines a dataset about the primary causes of obesity, with the goal of referencing excessive intake of caloric food , due to a lack of physical activity, a decrease in energy expenditure, genetics, socioeconomic variables, despair and anxiety. Datasets were gathered from the UCI Machine Learning Repository, while another dataset was gathered from several Tamil Nadu colleges. This research explores how machine learning algorithms such as Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Gradient Boosting, and Ada Boost work. The Support Vector Machine and Logistic Regression models are also tuned using Hyperparameter tuning techniques which includes Grid search and Randomized search. The Ensemble models includes Bagging, Ada Boost and Voting classifier are also implemented, in which the voting classifier combines all the models applied. to the datasets. The measures used for the prediction test are accuracy, precision, recalling and f1 score. According to the evaluation measures considered, the result shows that the Logistic Regression model has the best prediction accuracy followed by Decision tree and other models.**

Keywords-Obesity, Machine Learning , Gradient Boosting, Ada boost, Support vector, Logistic Regression, Decision Tree, Random Forest, Voting Classifier)

I. INTRODUCTION

The Excessive fat build-up in various body areas is a health risk that has a variety of effects on people of various ages and genders. The World Health Organization (WHO) claims that, everybody over the age of 18 is at risk of weight gain due to a variety of variables including calorie-dense foods, sedentary lifestyles, family history with obesity, water consumptions, and modes of transportation.

Obesity is a multi-factorial condition characterized by uncontrolled weight gain due to a high fat content, a high calorie intake, and insufficient energetic expenditure The capacity to forecast any obesity based on height, weight, age, and other characteristics can aid pediatricians in presenting meaningful information from a procedural standpoint. Obesity, on the other hand, is growing at an alarming rate, making exact prediction difficult. As a result, strategies for recognizing and decreasing growing obesity levels, such as obesity prediction and analysis, are crucial. In order to anticipate and prevent obesity, the Obesity Prediction technique employs algorithms to analyse large amounts of data.

Obesity prediction is the most often used approach, and it frequently analyses pre-existing obesity data to identify the height and weight of players who have engaged in risky behaviors and are at a high chance of becoming fat. Machine learning deal with statistical methods which is a part of Artificial intelligence and develops a system that is able to learn and adapt without following explicit instructions from the past experiences.

Machine learning techniques, data, and statistical algorithms are used in predictive analytics to determine the chance of future events, based on existing data. The purpose is to provide the best judgments of what will happen in the future, and rather than knowing what has happened. Predictive analysis entails the creation of models that aid in prediction. Gradient boosting, decision trees and classification, support vector machines, regression, and random forest are some of the techniques used. By identifying the crime patterns automatically, the cops can take relevant measures to prevent/stop them. If the tool is not used, it could take many weeks or years of refining through a database to discover a pattern, or it might be missed altogether.

978-1-6654-8035-2/22/$31.00 ©2022 IEEE

II. Literature Review

Xueqin Pang, et. Al (2021) have proposed a Expectation of youth weight with machine learning and electronic wellbeing record information. This review evaluates seven machine learning models made to foresee juvenile corpulence from the ages of 2 to 7 utilizing EHR information up to age 2 years [1].

B Singh, H Tawfik (2020) have proposed a Early Prediction of the risk of overweight and obesity in young people. Using machine learning many applications have developed like keratoconus detection and diabetes prediction [11][12]. The analysis makes use of a large quantity of data to build a machine learning-based model that can identify young individuals who are at risk of becoming overweight or obese [2].

Rodolfo Canas Cervantes, Ubaldo Martinez Palacio (2020) have proposed a machine learning approach on estimation of obesity levels on computational intelligence. The thesis uses data from three countries: Colombia, Peru and Mexico. The outcomes uncover a helpful instrument for doing a correlation study among the expressed strategies, including methods like Decision Tree, Support Vector Machine, and fundamental K-Means.

Gonzalo Colmenarejo (2020) have proposed Childhood and Adolescent Obesity prediction model using machine learning [3]. The field of Machine Learning models for foreseeing youth and young adult stoutness and related results is analyzed, with an attention on the latest models that join profound learning with EHR.

Eduardo De-La-Hoz-Correa, et. Al (2019) have suggested a Obesity Level Estimation Software based on Decision Trees [4]. To choose, investigate, and model the data set, the work used the SEMMA data mining approach, and three methods were chosen: Decision Tree, Bayesian Networks, and Logistic Regression.

Mathias J. Gerl, et, al (2019) have proposed a Machine Learning of human plasma lipidomes for heftiness assessment in a huge populace associate [5]. The scientists looked to foresee various stoutness pointers dependent on the plasma lipidome in an enormous populace companion utilizing refined machine learning methods. In a novel mass spectrometric shotgun procedure, the degrees of 183 plasma lipid species were measured in an aggregate of 1,061 individuals from the FINRISK 2012 populace companion.

Calabria-Sarmiento (2018) have proposed a Software Applications to Health Sector[6]. The work provides a systematic evaluation of the literature that gathers various breakthroughs that have led to the solution of various problems in the sector, as well as advancements that have contributed to the processes of continual detection and treatment improvement [7].

Craig A. Biwer (2017) have proposed a Computing Obesity: Signal Processing and Machine Learning Applied to Predictive Modelling of Clinical Weight-Loss [8]. Due to headways in clinical checking abilities and current sign handling draws near, just as advancement AI calculations, models for foreseeing a singular's chances of progress are becoming useful. By breaking down the development examples of overweight and large individuals, an example fosters that obviously recognizes the people who are probably going to get more fit from the people who are probably not going to shed pounds.

RG Hall (2017) have provided a Regulated Machine-Learning Reveals that Old and Obese People Achieve Low Dapsone Concentrations [9]. The populace pharmacokinetics of dapsone in overweight and large individuals in the United States is explored in this review, which thinks about covariate impacts.

Fabio Mendoza Palechor, et. Al (2016) have proposed a Supervised and Unsupervised Data Mining Techniques on Cardiovascular Disease Analysis Using. Among the systems utilized in the proposed technique for the examination of sicknesses, including cardiovascular diseases, are choice trees, support vector machines, bayesian organizations, and k-closest neighbors [10].

III. Methodology

A good prediction technique speeds up the evolution of obesity data sets, aids in obesity prediction, and keeps track of resources related to obesity analysis. A fitting data mining methodology, machine learning techniques, and statistical tools can all be used to analyse obesity. Gradient boosting and support vector machine are two further strategies for obesity prediction and analysis described in this work. Logistic Regressions, Ada Boost, Random Forest, Decision Tree, and SVM were used to create this system.

The fundamental goal of this system is to try and predict the rate of obesity so that risk factors can be identified and preventative measures can be done by pediatrics or individuals. Obesity prediction is crucial for society's obesity prevention since it assists pediatrics in developing effective solutions. Obesity reduction will benefit individuals in a variety of ways. As a result, by forecasting obesity using various machine learning algorithms, obesity rates can be reduced

A. Dataset Description

TABLE I Description of Datasets

Dataset name	Dataset Source	No. of Instances	No of Attributes
Dataset 1	UCI machine learning repository, Github.	2711	17
Dataset 2	Realistic dataset collected from various college students from Tamil Nadu.	553	16

The description of Table 1 is given above, where the dataset's properties comprise the types of components that include Gender, age, height, weight, and family history of obesity are all factors to consider. Consumption of high-calorie foods on a regular basis, consumption of vegetables on a regular basis, number of main meals per day, and water consumption, Tobacco use, Calorie counting, Time spent on digital devices each day (e.g., laptops or cell phones), alcohol consumption.

B. *Proposed System Architectures*

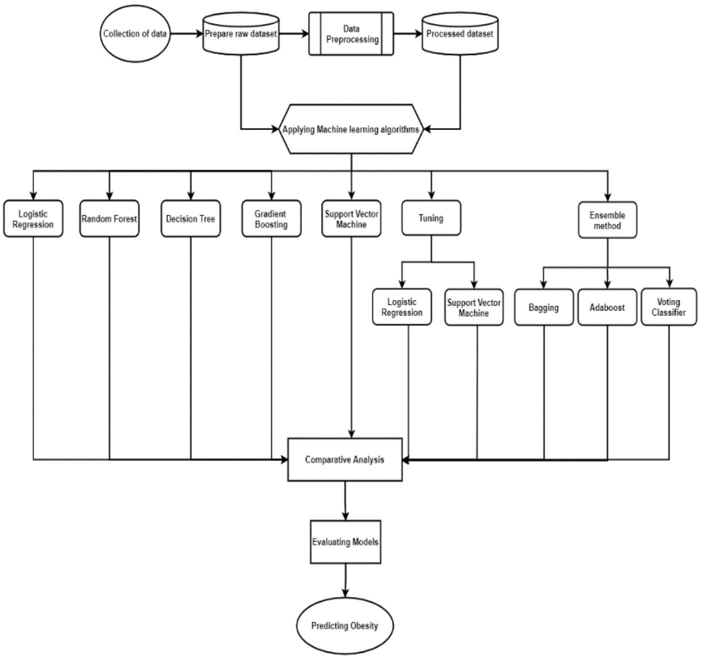

Figure. 1 Proposed System Architecture

Fig. 1 shows the flow diagram of the proposed system. The Obesity datasets are collected and processed for Scaling the dataset to get the normalized values and separating the dataset into training and test sets to train the Model. Various Algorithms are used and the algorithm which provides best accuracy is chosen

C. *Data Pre-Processing*

To normalize the values in the dataset, a scaling technique is used. Scaling is an important pre-processing technique that is needed to standardize/normalize the input data, with scaled values ranging from 0 to 1. When the range of values in each column is substantially different, they must be scaled to a similar level. After bringing the numeric values to a common level, the input data can be subjected to a machine learning algorithm.

D. *Model Building*

The proposed research work uses various machine learning models for Obesity prediction and such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine and Ada Boost. The Support Vector Machine and Logistic Regression models are an also tuned using Hyper parameter tuning technique which includes Grid search and Randomized search. The Ensemble models includes Bagging, Ada Boost and Voting classifier are also implemented, in which the voting classifier combines all the models applied. to the datasets.

E. *Classification Report*

The Following are the Prediction measures that are used for model evaluation described in the proposed system,

i. Precision,
ii. Recall,
iii. F1 Score.

Precision:

Precision is the ratio of positively predicted observations to the total positive predicted observations.

Precision: - Accuracy of positive predictions.

$$Precision \ = \frac{TP}{TP+FP} \qquad \textbf{(1)}$$

The above equation 1 gives the ability of the model to predict the instance positive that is actually positive.

Recall:

Recall is the ratio of positively predicted observations to the total positively predicted observations.

Recall:- Fraction of positives that were correctly identified.

$$recall \ = \frac{TP}{TP+FN} \qquad \textbf{(2)}$$

The above equation 2 gives the ability of the model to predict all the positive instances.

F1 score:

F1 is a measure of a accuracy on a dataset. The F1 score used for evaluating binary classifications systems. The F1 score will classify examples into positive or negative.

$$f1\ score = \frac{2*Precision*Recall}{Precision\ +Recall} \qquad \textbf{(3)}$$

The above equation 3 measures the accuracy with weighted average using the precision and recall measures.

IV. Results and Discussion

The effectiveness of our suggested methodology is evaluated using metrics obtained from the experiments, and the fine-tuning technique of machine learning is explored. The experimental study is based on two datasets: one that is publicly available and the other that has been collected from various universities in Tamil Nadu through web.

Experiments are run on both datasets using machine learning algorithms: Ada boost, Decision Tree, Random Forest, Logistic regression, Support Vector Machine (SVM), and Gradient Boost after Scaling. The experiments are carried out with and without the weight attribute since the weight attribute is highly correlated with the target class.

TABLE II Predictive accuracy of Proposed models with weight

Models	Training &Testing		Cross validation	
	Dataset1	***Dataset2***	***Dataset1***	***Dataset2***
LR	98.58	85.54	95.55	86.43
DT	99.05	93.97	99.29	97.28
RF	99.36	94.57	98.86	96.02
GB	99.05	94.57	98.86	96.20
AB	98.51	94.57	98.95	96.02
SVM	98.51	84.93	97.06	84.81

From Table II The Random Forest (RF) achieves a highest accuracy of 99.36 % on the dataset 1 and dataset 2 having highest accuracy of 94.57 for RF, GB and AB.

A. Tuning for LR & SVM

The Table III and Table IV represent the obtained results of LR and SVM with tuning supported by Randomized search and Grid search. Randomized search provides improved results for SVM and Grid search provides improved results for LR.

TABLE III Tuning for Dataset 1

Models	Randomized Search	Grid Search
LR	97.63	99.68
SVM	99.21	98.89

TABLE IV Ensemble Methods for Proposed models

Models	Dataset1	Dataset2
Voting classifier	99.48	96.92
Bagging	99.41	96.01

Table IV gives the complete prediction measure results as the voting classifier provides 99.48% accuracy for dataset1 and 96.92% accuracy for dataset2. Bagging models supports with accuracy of 99.41% for dataset1 and 96.01% for dataset2.

From the fig.2 the best AUC value supported to the LR after tuning with grid search and next highest AUC value is supported for GB.

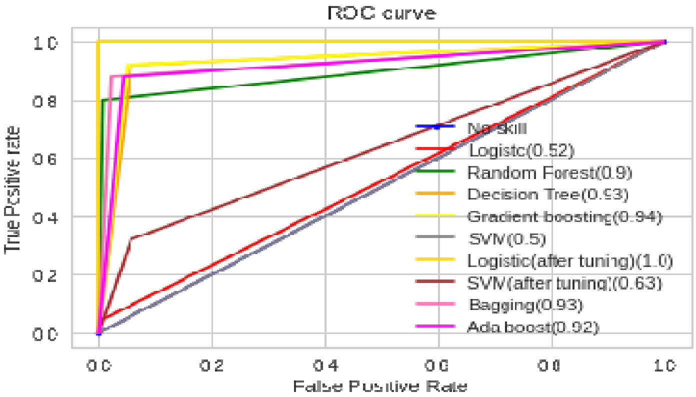

Figure. 2 Results of ROC curve

Fig. 3 represents the results of proposed models along with various measures. The outstanding results expressed for the Random forest and other models obtained the results above 98%.

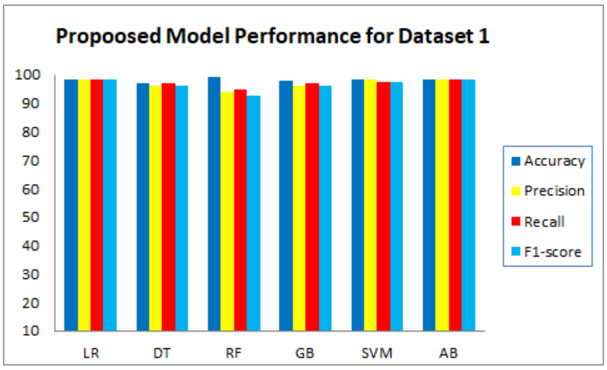

Figure. 3 Results of proposed models for Dataset1

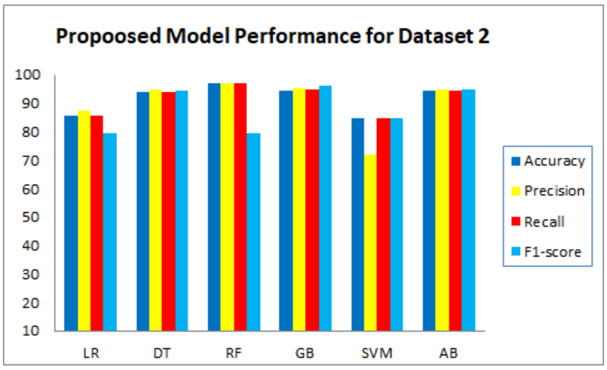

Figure. 4 Results of proposed models for Dataset2

Dataset2 results of proposed models have represented in figure 4. Decision tree provides best results for accuracy, precision and recall. Boosting methods (AB, GB) obtained the better results for dataset2.

V. CONCLUSION

In the growing research advancement, predicting the obesity on basis of machine learning models aims at reducing the obesity rate. Machine learning models like Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, and Gradient Boosting are utilized in this recommended framework and Tuning of some fundamental machine learning models and Ensemble procedures have been utilized to faster a methodology dependent on computational insight. Since the weight characteristic is substantially connected with the target class, the studies were performed with and without the attribute. Before tuning the parameters the LR model achieves the best accuracy of 99.36% for dataset1 and dataset 2 obtained good results with decision tree.

The Logistic Regression accomplishes a forecast exactness of 99.68 % and SVM with 99.21% after performing tuning. Ensemble methods have been supporting 99.48% with voting classifier and 99.41% with bagging methods. The proposed work further has to develop a mere optimized model using deep learning.

REFERENCES

[1] Pang, X., Forrest, C.B., Lê-Scherban, F. and Masino, A.J., 2021. Prediction of early childhood obesity with machine learning and electronic health record data. International Journal of Medical Informatics, 150, p.104454.

[2] Singh, B. and Tawfik, H., 2020, June. Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People. In International Conference on Computational Science (pp. 523-535). Springer, Cham.

[3] Cervantes, R.C. and Palacio, U.M., 2020. Estimation of obesity levels based on computational intelligence. Informatics in Medicine Unlocked, 21, p.100472.

[4] Colmenarejo, G., 2020. Machine learning models to predict childhood and adolescent obesity: A review. Nutrients, 12(8), p.2466.

[5] De-La-Hoz-Correa, E., Mendoza Palechor, F., De-La-Hoz-Manotas, A., Morales Ortega, R. and Sánchez Hernández, A.B., 2019. Obesity level estimation software based on decision Trees.

[6] Gerl, M.J., Klose, C., Surma, M.A., Fernandez, C., Melander, O., Männistö, S., Borodulin, K., Havulinna, A.S., Salomaa, V., Ikonen, E. and Cannistraci, C.V., 2019. Machine learning of human plasma lipidomes for obesity estimation in a large population cohort. PLoS biology, 17(10), p.e3000443.

[7] Calabria-Sarmiento, J.C., Ariza-Colpas, P., Pineres-Melo, M., Ayala-Mantilla, C., Urina-Triana, M., Morales-Ortega, R., Peluffo-Martinez, G., Mendoza-Palechor, F. and Echeverri-Ocampo, I., 2018. Software applications to health sector: a systematic review of literature.

[8] Biwer, C., 2017. Computing Obesity: Signal Processing and Machine Learning Applied to Predictive Modeling of Clinical Weight-Loss (Doctoral dissertation).

[9] Hall, R.G., Pasipanodya, J.G., Swancutt, M.A., Meek, C., Leff, R. and Gumbo, T., 2017. Supervised Machine - Learning Reveals That Old and Obese People Achieve Low Dapsone Concentrations. CPT: pharmacometrics & systems pharmacology, 6(8), pp.552-559.

[10] Palechor, F.M., De la Hoz Manotas, A., Colpas, P.A., Ojeda, J.S., Ortega, R.M. and Melo, M.P., 2017. Cardiovascular Disease Analysis Using Supervised and Unsupervised Data Mining Techniques. J. Softw., 12(2), pp.81-9.

[11] Nirmaladevi, K., Shanthi, S., Agila, T., Dharani, R. T., & Dhivyapriya, P. (2020). Analysis and prediction of diabetes using machine learning. Test Engineering and management, Vol.83, No.9, pp.14533-14538.

[12] Shanthi, S., Nirmaladevi, K., Pyingkodi, M., Dharanesh, K., Gowthaman, T., & Harsavardan, B. (2021, February). Machine learning approach for detection of keratoconus. In IOP Conference Series: Materials Science and Engineering (Vol. 1055, No. 1, p. 012112). IOP Publishing