International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 10 No. 6S, 2024
https://www.theaspd.com/ijes.php

Health Care Technologies and Analytics: Transforming Modern Healthcare

Hirenkumar jesalbhai vasava[1], Priteshkumar B Vasava[2], Yoginkumar G Garasia[3], Mitulkumar dahyabhai prajapati[4], Jatinkumar Chaudhari[5], Priyanka Sumantrai Patel[6]

[1]Email: hiren.vasava@gmail.com
Government Engineering College, Bharuch, Gujarat, India.
[2]Email: priteshvasava7187@gmail.com
Government Engineering College, Bharuch, Gujarat, India.
[3]Email: yogingarasia@gmail.com
Government Engineering College, Bharuch, Gujarat, India.
[4]Email: mdpgecv@gmail.com
Government Engineering College, Bharuch, Gujarat, India.
[5]Email: jatinchaudhari@hotmail.com
Government Engineering College, Bharuch, Gujarat, India.
[6]Email: priyankapatelgec@gmail.com
Government Engineering College, Bharuch, Gujarat, India.

Abstract: Obesity is a growing global health problem that increases vulnerability to many chronic diseases. Several chronic diseases are attributed to obesity including diabetes and cardiovascular disorders. Machine learning methods have potential in predicting the risk of obesity using lifestyle and physiological parameters. There is a need for efficient and accurate methods to predict the risk of obesity. This paper compares logistic regression, decision tree, random forest, support vector machine (SVM), and gradient boosting methods using a voting classifier to predict the risk of obesity. A healthcare-related dataset was used in this study and the performance was evaluated using accuracy and classification metrics. The results show that using ensemble learning models improve the prediction performance especially in the case of using multiple models. This paper contributes to justifying the use of ensemble methods in the field of healthcare analytics. Gradient Boosting
(LightGBM) had the best performance, and thus it is the most suitable model for this task. Gradient Boosting (LightGBM) had the best performance, and thus it is the
most suitable model for this task. The technique can be adopted to develop efficient and reliable decision-making applications in healthcare.

INTRODUCTION:

Obesity is a global crisis of public health, inflicting misery on enormous populations and risking numerous chronic diseases, from which it reduces an individual's life expectancy by 10 years. Thus, it becomes critical to identify these persons early on so that intervention and prevention strategies may be effectively applied to them [1]. Conventionally, obesity risk assessment has relied on clinical evaluations and statistical models, with Body Mass Index (BMI) as the standard measure. The BMI is defined as weight in kilograms divided by height in meters squared (kg/m2) and is categorized as underweight (BMI < 18.5), normal (BMI 18.5-24.9), overweight (BMI 25-29.9), and obesity (BMI ≥30), which is further divided into three classes. Waist circumference is another important parameter that has sometimes shown better correlations with cardiovascular disease risk, where a waist circumference greater than 102 cm (40 inches) for men and 88 cm (35 inches) for women, measured at the level of the iliac crest, constitutes obesity [2].

Obesity is a complicated condition and is influenced by genetic, behavioral, and neuroendocrine factors; it is generally due to an imbalance in energy intake and expenditure, resulting in the excessive accumulation of fat [3]. The prominent behavioral risk factors associated with obesity are lack of exercise, poor diet-selection in particular for the consumption of ultra-processed foods, comforting oneself by eating after stopping smoking, binge eating, and night eating syndrome [4]. The other conditions that contribute to obesity are hypothalamic dysfunction, growth hormone deficiency, hypothyroidism and hypogonadism, and polycystic ovarian syndrome (PCOS) and Cushing's syndrome. In addition, some drugs increase the risk of weight gain and thus pose a further risk of exacerbating obesity: these include corticosteroids, antipsychotics, certain antidepressants, beta blockers, and some antidiabetic medications such as insulin and sulfonylureas.

Characteristically, obesity has several health complications, affecting several organ systems. Cardiovascular include hypertension, hyperlipidemia, heart disease, stroke, and metabolic syndrome, with morbidity and mortality directly dependent on these diseases. Obstructive sleep apnea and obesity hypoventilation syndrome are some respiratory complications affecting general health [5]. Gastrointestinal problems with acidic reflux diseases (GERD) and fatty liver disease can also be found among obese individuals. Major endocrine and metabolic disturbances such as type 2 diabetes, insulin resistance, and metabolic syndrome are also prominently related to extraordinary weight gain [6]. The above parameters influence the musculoskeletal system by increasing load on joints and causing osteoarthritis with a higher risk rate for gout. Urinary incontinence, venous thrombosis, hernias, and increased risk for a number of cancers, among others, are also consequences of obesity, thus calling for effective prevention and management strategies [7].

Managing obesity requires a multifaceted approach that addresses underlying causes and promotes sustainable lifestyle modifications. Effective weight management begins with lifestyle changes, including setting realistic goals such as a 5% reduction in body weight per year, following structured diet plans like low-calorie diets, and engaging in regular physical activity [8]. When lifestyle interventions are insufficient, pharmacotherapy can be considered. GLP-1 agonists help regulate appetite by slowing gastric emptying and enhancing insulin secretion, though they may cause nausea and vomiting [9]. Orlistat, another medication, inhibits pancreatic lipase to prevent fat absorption but can lead to gastrointestinal discomfort. Sympathomimetics, such as phentermine, enhance sympathetic nervous system activity, increasing energy expenditure and aiding in weight loss. These treatment approaches, when appropriately combined, can significantly improve obesity outcomes and reduce associated health risks [10].

Obesity is a significant health issue that is associated with many chronic diseases, as shown in figure 1. It has a high risk of causing conditions like diabetes, high cholesterol, high blood pressure, and kidney disease. Obesity is also linked to severe cardiovascular diseases, such as heart disease and stroke, liver disease, and dementia. Management of obesity via lifestyle changes, such as a proper diet and exercise, is essential in preventing these life-threatening illnesses and enhancing general well-being

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 10 No. 6S, 2024
https://www.theaspd.com/ijes.php

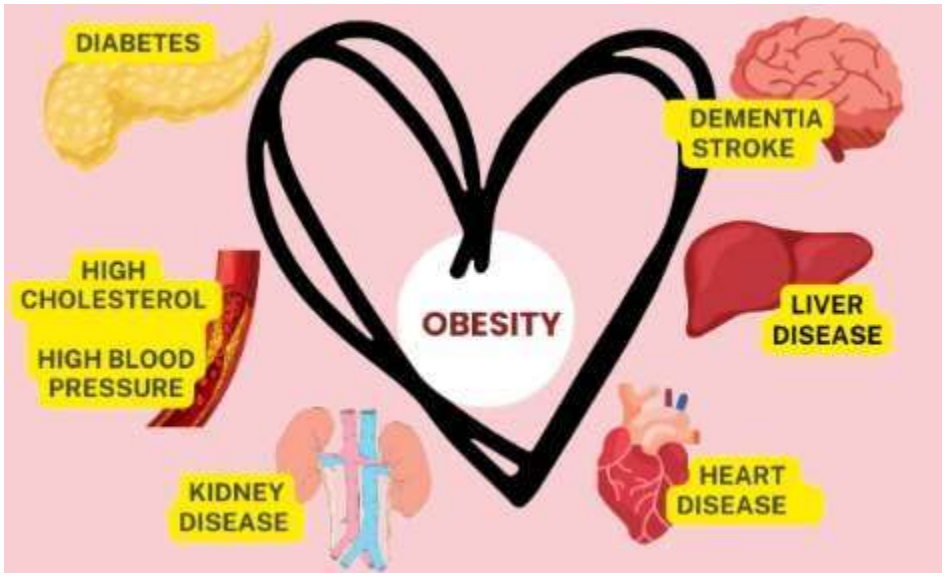

Figure 1: Overview of disease caused by Obesity

Related work: Obesity has become a significant public health concern worldwide, with its prevalence increasing at an alarming rate. Various research studies have been conducted to predict obesity risk using machine learning techniques, aiming to enhance early diagnosis and intervention. The application of machine learning provides valuable insights by analyzing diverse factors such as demographics, lifestyle habits, and medical history. Several studies have explored different methodologies, algorithms, and datasets to improve predictive accuracy and identify crucial obesity-related risk factors.

Ferdowsy et al. (2021) applied nine machine learning algorithms to predict obesity risk based on a dataset of over 1,100 individuals. The study found that Logistic Regression achieved the highest accuracy of 97.09%, while Gradient Boosting performed the poorest. The advantage of this study is its high predictive accuracy, but its limitation lies in dataset constraints, as it is focused primarily on the Bangladeshi population. Similarly, Dugan et al. (2015) used six machine learning models, analyzing data from a clinical decision support system called CHICA to predict childhood obesity. The ID3 model achieved the best performance with an 85% accuracy rate. While the study highlights the importance of early childhood data in obesity prediction, it is limited to data collected before the second birthday.

Zheng and Ruggiero (2017) explored obesity prediction in high school students using four enhanced machine learning models: binary logistic regression, improved decision tree (IDT), weighted K-nearest neighbor (KNN), and artificial neural networks (ANN). The weighted KNN model achieved the highest accuracy (88.82%), followed by ANN (84.22%) and IDT (80.23%). This study's strength is its focus on adolescent obesity risk factors, but its findings are specific to a single dataset from Tennessee. Dirik (2023) conducted a comparative study using multiple machine learning techniques, including Random Forest, Logistic Regression, and Decision Tree, to predict obesity. The study found that Random Forest performed best, with an accuracy of 95.78%. Although the research provides a strong comparative analysis, its limitation is the lack of generalizability to diverse populations.

Rodríguez et al. (2021) investigated the effectiveness of machine learning techniques in predicting obesity and overweight conditions. The study found that Random Forest was the best-performing model, achieving 78% accuracy. This research underscores the potential of machine learning for obesity identification but is limited by its reliance on self-reported eating habits and physical

condition data. Similarly, Pang et al. (2021) developed seven machine learning models using electronic health record (EHR) data from 860,510 patients. Their study demonstrated that XGBoost was the most effective model, with an area under the curve (AUC) score of 0.81. The strength of this research lies in its large dataset, but its limitation is the dependency on high-quality EHR data.

Jeon et al. (2023) examined age-specific risk factors for obesity using machine learning classifiers. The study identified triglycerides, ALT (SGPT), glycated hemoglobin, and uric acid as the most significant predictors of obesity. The study demonstrated over 70% accuracy for individuals aged 19–39 but observed a decline in accuracy for older age groups. The research highlights the role of age and gender in obesity prediction, but its limitation lies in the dataset's geographical constraints. Gerl et al. (2019) used machine learning to analyze human plasma lipidomes for obesity estimation. Their findings suggested that the lipidome is a strong predictor of body fat percentage ($R2 = 0.73$), surpassing traditional BMI measurements. The study emphasizes the significance of lipid profiling in obesity prediction, but its findings require further validation in clinical settings.

Maulana et al. (2024) developed an obesity prediction model using the CatBoost algorithm, achieving an accuracy of 95.98%. The study identified weight, height, gender, dietary habits, and physical activity as the most influential factors in obesity prediction. Although the CatBoost model demonstrated superior performance, its reliance on a dataset from Latin America limits its global applicability. Peng et al. (2025) investigated obesity risk factors among older adults in China using machine learning and SHapley Additive exPlanations (SHAP) for model interpretation. Their findings highlighted gender, transportation-related physical activity, and road network density as crucial determinants of obesity. The study underscores the environmental and behavioral contributors to obesity, but its limitation is the relatively small sample size.

Overall, these studies demonstrate the growing importance of machine learning in obesity prediction and risk assessment. Various algorithms, including Logistic Regression, Random Forest, XGBoost, and CatBoost, have shown promising results in different populations and datasets. However, challenges such as dataset limitations, model generalizability, and reliance on specific features remain critical areas for future research. Integrating machine learning with real-time health monitoring and diverse demographic datasets can further enhance obesity prediction models, aiding in early intervention and preventive healthcare strategies.

Table 1: Previous Related work to Obesity in domain of AI

Reference	Objective	Methodology	Advantage	Limitations
[11] Ferdowsy et al. (2021)	Develop an ML-based model for obesity risk prediction.	Collected data from 1100+ individuals and applied nine ML algorithms (k-NN, RF, LR, MLP, SVM, Naïve Bayes, AdaBoost, DT, Gradient Boosting).	Logistic Regression achieved the highest accuracy (97.09%). Identifies obesity risk and contributing factors.	Gradient Boosting had the lowest accuracy (64.08%). Dataset limited to a specific population.

[12] Dugan et al. (2015)	Predict childhood obesity using clinical decision support data.	Analyzed six ML models (RandomTree, RandomForest, J48, ID3, Naïve Bayes, Bayes) on CHICA dataset.	ID3 model achieved 85% accuracy and 89% sensitivity. Provides insights into childhood obesity predictors.	Focused only on early childhood obesity; dataset restricted to CHICA system.
[13] Zheng & Ruggiero (2017)	Predict obesity in high school students using ML models.	Used binary logistic regression, improved decision tree (IDT), weighted k-NN, ANN on YRBSS dataset.	Weighted k-NN achieved 88.82% accuracy, ANN achieved 84.22%. Considers risk and protective factors.	Logistic Regression performed poorly (56.02% accuracy). Limited to Tennessee YRBSS data.
[14] Dirik (2023)	Compare ML techniques for obesity prediction.	Evaluated ML models (MLP, SVM, FuzzyNN, FURIA, RS, RT, RF, NB, LR, DT) on obesity-related data.	RF model had highest accuracy (95.78%), LR (95.22%). Provides multiple evaluation metrics.	Model performance varies based on dataset. No real-world deployment discussed.
[15] Rodríguez et al. (2021)	Develop an ML model to predict overweight/obesity.	Applied Decision Tree, SVM, k-NN, Naïve Bayes, MLP, RF, Gradient Boosting, XGBoost.	RF performed best (78% accuracy, 79% precision). Demonstrates practical applications in healthcare.	Model accuracy is moderate compared to other studies. Feature importance analysis not detailed.
[16] Pang et al. (2021)	Predict childhood obesity using EHR data.	Used EHR data from 860,510 patients, applied seven ML models, including XGBoost.	XGBoost had the highest performance (AUC 0.81). Generalizes to both genders.	Requires high-quality EHR data. Complex preprocessing of clinical data.
[17] Jeon et al. (2023)	Identify age-specific obesity risk factors using ML.	Assessed six ML models on KNHANES data for 21,100 participants.	Highlights age- and gender-specific risk factors. Achieved >70% accuracy for some age groups.	Accuracy varies by age and gender. Feature selection impact on accuracy not fully explored.

[18] Gerl et al. (2019)	Predict obesity using plasma lipidomics and ML.	Used lipidomic data from FINRISK 2012, applied multiple ML models.	Lipidome-based obesity prediction outperformed BMI-based methods. Identifies molecular markers.	Requires specialized lipidomic data. Limited generalizability outside studied cohort.
[19] Maulana et al. (2024)	Develop an obesity prediction model using CatBoost.	Used demographic, lifestyle, and health-related features from 2,111 individuals; compared CatBoost to other ML models.	CatBoost outperformed other models (95.98% accuracy). Feature importance analysis aids prevention strategies.	Requires further validation on diverse populations. Limited feature diversity.
[20] Peng et al. (2025)	Identify overweight/obesity risk factors in older adults using ML.	Survey data from 400 older adults, applied six ML algorithms, used SHAP for interpretation.	CatBoost performed best. Identified key risk factors (gender, physical activity, road network density).	Small dataset (400 participants). Focused only on older adults in China.

2. METHODOLOGY

Data Collection and Preprocessing The dataset comprises various lifestyle, demographic, and physiological variables related to obesity risk. This dataset appears to be related to obesity classification based on various demographic, lifestyle, and dietary factors. Here's a brief breakdown of the attributes:

- Gender: Male/Female
- Age: Continuous numerical value
- Height: In meters
- Weight: In kilograms
- family_history_with_overweight: Yes/No (Indicates if obesity runs in the family)
- FAVC (Frequent consumption of high-calorie food): Yes/No
- FCVC (Frequency of consumption of vegetables): Numerical (Likely on a scale)
- NCP (Number of main meals per day): Numerical
- CAEC (Consumption of food between meals): Categories like "Sometimes," "Frequently," etc.
- SMOKE: Yes/No (Smoking habit)

- CH2O (Daily water intake in liters): Continuous numerical value
- SCC (Calories consumption monitoring): Yes/No
- FAF (Physical activity frequency in hours per week): Continuous numerical value
- TUE (Time using technology in hours per day): Continuous numerical value
- CALC (Alcohol consumption frequency): Categories like "Sometimes," "Frequently," etc.
- MTRANS (Mode of transportation): Categories like "Public_Transportation," "Automobile," etc.
- NObeyesdad (Obesity classification): Categories such as "Normal_Weight," "Obesity_Type_III," "Overweight_Level_II,"

Exploratory Data Analysis (EDA) Visual EDA techniques are applied to understand the distribution of obesity-related factors, correlations among variables, and feature importance. The Figure 2 provided visualization reveals information about the distribution of the levels of obesity among people in the dataset. The most common category is Obesity_Type_III, with 4,046 instances, then Obesity_Type_II (3,248), and Normal_Weight (3,082). The dataset also includes a high number of people who are categorized as Obesity_Type_I (2,910) and Insufficient_Weight (2,523). Overweight_Level_I and Overweight_Level_II also appear in similar numbers, at 2,427 and 2,522 people, respectively.

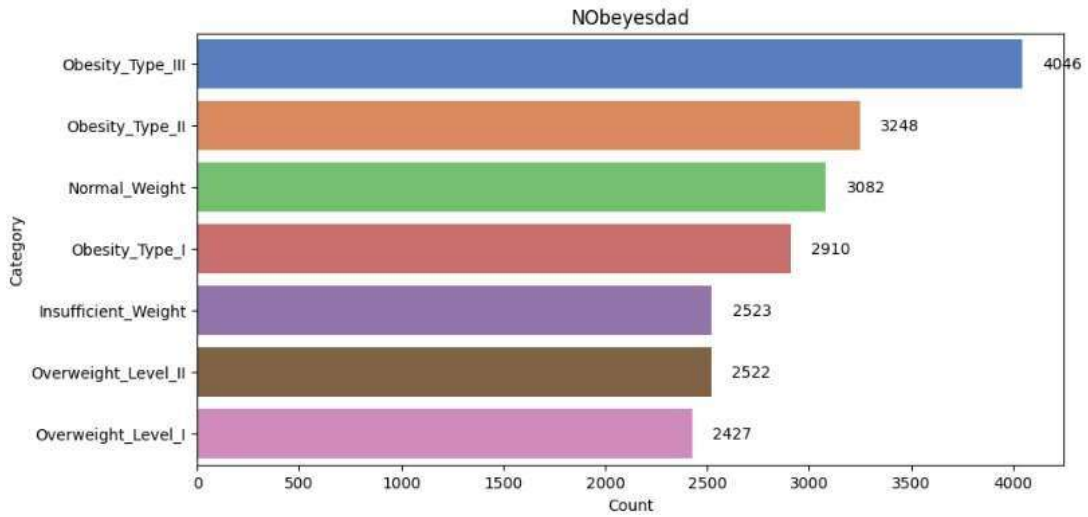

Figure 2: Overview of different Category and its count

The Figure 3 shows the gender split across the dataset. The population is almost evenly balanced, with 50.2% Female and 49.8% Male. This level of evenness ensures that any gender-based findings derived from the data are unbiased and representative in nature. A dataset such as this one with a balanced proportion is good for machine learning algorithms since it decreases the likelihood of gender-based prejudice in predictive modeling. This spread also suggests that any health-related trends found within the dataset are not biased toward one gender, so the findings are more applicable.

Figure 3: Gender Distribution

The Figure 4 shows the distribution of those with and without a family history of overweight. Most, 82.0%, have a family history of overweight, compared to just 18.0% who do not. This indicates a strong environmental or genetic factor in weight-related conditions. The high percentage of those with a family history of overweight could imply hereditary factors or common lifestyle factors leading to weight gain. This knowledge is important for identification of the risk factors for obesity and may aid in the development of interventions specifically for those at increased risk.

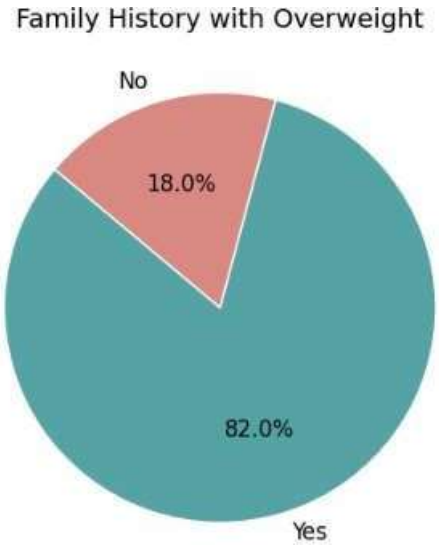

Figure 4: Overview of family history with overweight

Figure 5 is a comparative distribution of numerical features between the original dataset and the training dataset. Each subplot corresponds to a different numerical variable, with the density distributions for both datasets shown. The blue shaded regions are the training data distribution, and the red overlay is the original dataset distribution.

The most important point from this visualization is that the distributions of the training data closely follow those of the original dataset for the majority of features, implying that the training

set well represents the original population. The minor differences in some variables, like height, weight, and certain health-related measures show minor differences in feature representation. The values of train divergences offer a quantitative estimate of the degree to which the distributions differ. Lower values of divergence indicate closer alignment, and higher values are indicative of possible shifts in feature distributions. Recognizing such differences is of prime importance in ensuring that the model generalizes and is unbiased toward certain subsets of the data. If sizeable differences were found, then other data preprocessing methods, such as resampling or feature scaling, may have to be undertaken to improve the performance of the model.

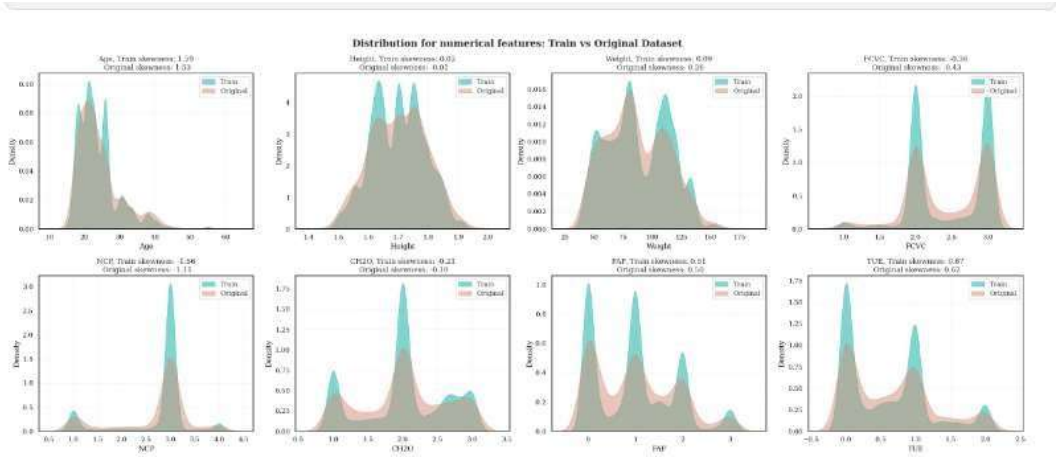

Figure 5: Distribution for numerical features: train vs original dataset

The Figure 6 shows the mean Body Mass Index (BMI) values by weight categories. The information is grouped into several levels of obesity, overweight, normal weight, and underweight. Obesity Type III has the highest mean BMI of 41.78, followed by Obesity Type II at 36.52 and Obesity Type I at 32.15. Overweight Level II and Overweight Level I have mean BMI of 28.19 and 26.06, respectively. Normal Weight individuals have a mean BMI of 22.00, whereas those in the category of Insufficient Weight possess the lowest mean BMI of 17.58.

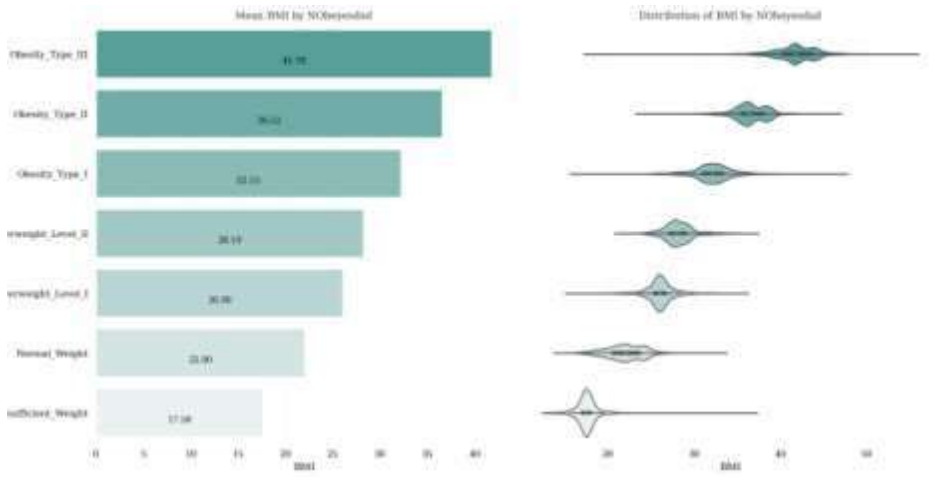

Figure 6: Mean BMI value of different category

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 10 No. 6S, 2024
https://www.theaspd.com/ijes.php

2.3 Model Implementation: This research adopts a systematic process of machine learning model building, such as data pre-processing, feature selection, model training, testing, and evaluation. The Figure 7 provides Block Diagram of proposed Model. The detailed steps are described as follows:

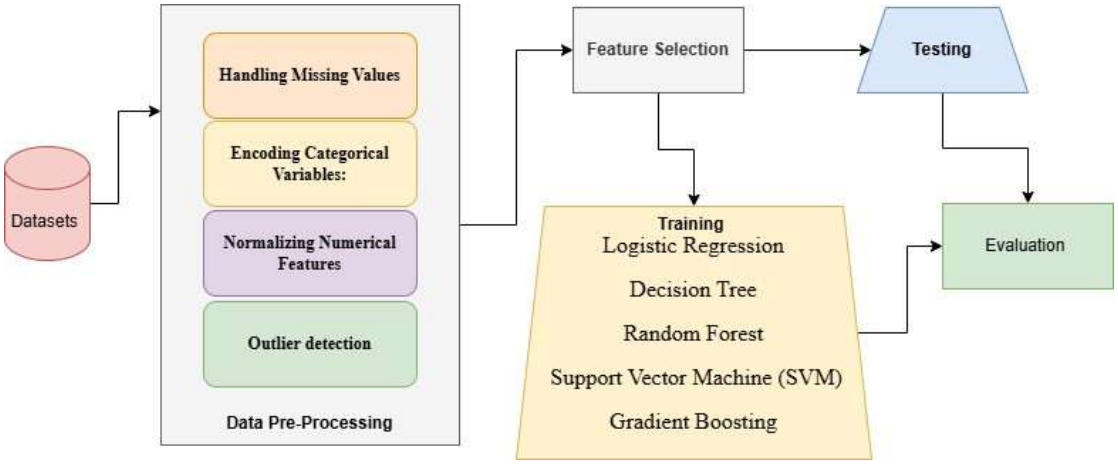

Figure 7: Block Diagram of proposed Model

Preprocessing steps include

- **Handling Missing Values:** Handling missing data is crucial for maintaining the quality and reliability of a dataset. If the number of missing values is minimal and does not significantly impact the dataset, rows or columns with missing values can be removed. No missing value found in the dataset.
- **Detection of Outliers:** Outliers are data points that significantly differ from the majority of observations. Detecting and handling them is crucial for improving model performance.Z-Score (Standard Score) method is use to detecte the outliers, which is based on standard deviations from the mean:

$$Z = \frac{X - \mu}{\sigma}$$

Common threshold: $|Z| > 3$ (2)

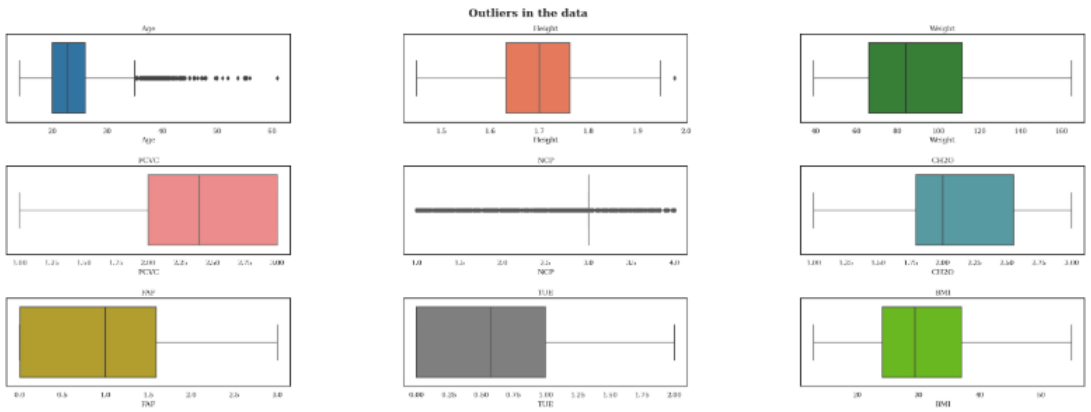

Figure 8: Outliers in the datasets

- **Encoding Categorical Variables:** Since machine learning algorithms work best with numerical data, categorical variables need to be encoded into numerical format. **One-Hot Encoding** method converts categorical variables into binary columns for each category. It is suitable for categorical variables with a small number of unique values.
- **Normalizing Numerical Features:** Normalization is essential to bring numerical features into a standard range, improving model performance and convergence. Min-Max Scaling is use to transforms features into a range between 0 and 1 using the formula

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

.

2. Feature Selection

Feature selection is performed to identify the most relevant attributes that contribute to the prediction task. Techniques such as correlation analysis, Recursive Feature Elimination (RFE), and tree-based importance methods are used to reduce dimensionality and improve model performance.

3. Model Training

Several machine learning models are trained to classify the target variable effectively. The models considered include:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Gradient Boosting (LightGBM)

Hyperparameter tuning is performed to optimize each model using Grid Search techniques

4. Testing

The dataset is split into training and testing subsets to evaluate model performance. The test set is used to assess the generalization ability of the trained models.

5. Evaluation

The performance of the trained models is evaluated using various metrics such as:

Accuracy – Measures overall correctness.

Precision, Recall, and F1-Score – Evaluates classification performance, especially for imbalanced datasets.

3. **RESULTS AND DISCUSSION**

The analysis reveals significant relationships between lifestyle factors and obesity risk. Feature importance analysis indicates that factors such as dietary habits, physical activity levels, and BMI play a crucial role in predicting obesity risk. The efficiency of different machine learning models was ascertained through Accuracy, Precision, Recall, and F1-Score to ascertain their classification

strength. The results shown in table 2 and figure 9 provides that Gradient Boosting (LightGBM) recorded a maximum accuracy of 94%, while Random Forest recorded 92%, which is ascertained to be high in terms of predictive strength. Logistic Regression (67%) was the least accurate, implying that it may not be able to capture intricate patterns. Although its F1-score (0.7306) implies a decent precision-recall balance, it falls behind other models. Support Vector Machine (SVM) (76%) had a significant improvement over Logistic Regression, with an improved precision (0.7552) and recall (0.749) balance, which makes it a better model for classification tasks. Decision Tree (88%) was much better with high recall (0.843), which implies that it properly classified positive cases but with comparatively low precision (0.8236), reflecting some misclassification. Random Forest (92%) was better than Decision Tree in performance, depicting good precision (0.8905) and recall (0.892) in reducing overfitting as compared to an individual tree model. Gradient Boosting (LightGBM) (94%) performed better than the rest of the models, having the highest precision (0.9605), recall (0.962), and F1-score (0.9006), which reflect better classification performance. This indicates that LightGBM is good at capturing intricate patterns with a high generalization ability.

Table 2: Result of different models before feature selection

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.67	0.6519	0.647	0.7306
Support Vector Machine (SVM)	0.76	0.7552	0.749	0.7308
Decision Tree	0.88	0.8236	0.843	0.8114
Random Forest	0.92	0.8905	0.892	0.8206
Gradient Boosting (LightGBM)	0.94	0.9605	0.962	0.9006

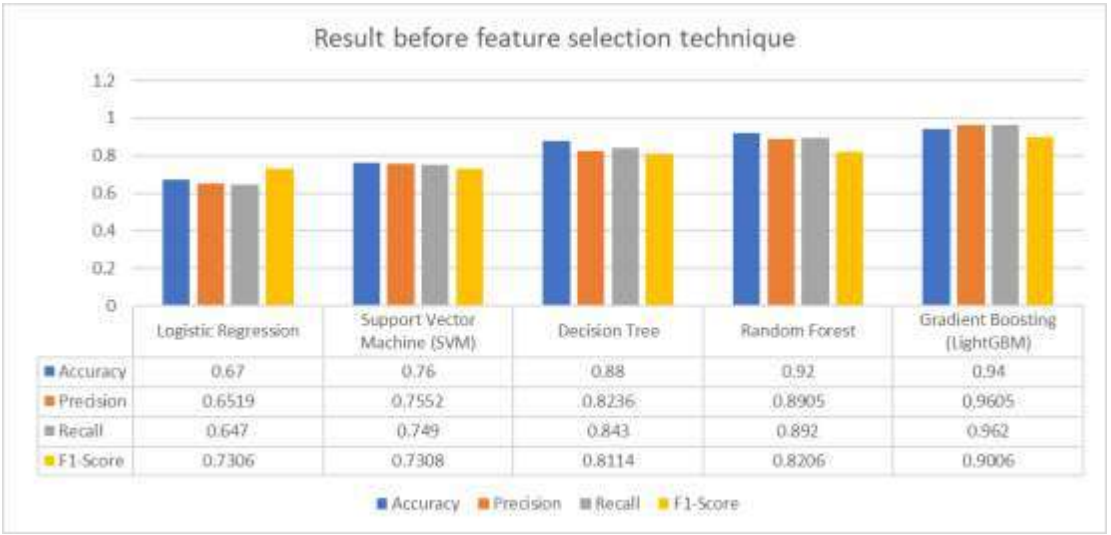

Figure 9: Result of different models before feature selection

The feature selection process was conducted to identify the most influential variables in predicting the target outcome. The **feature importance plot** above highlights the relative contribution of each feature, determined using **Gradient Boosting (LightGBM)**, which evaluates how each variable impacts the model's decision-making process.

Key observations from the analysis:

Weight, Height, and Age emerged as the most critical features, indicating their strong correlation with the prediction task.

Physical activity-related variables such as **FAF (Physical Activity Frequency) and TUE (Time using Technology)** also played a significant role, suggesting lifestyle factors influence the target outcome.

Nutritional habits like **CH2O (Water Intake), FCVC (Frequency of Vegetable Consumption), and NCP (Number of Meals per Day)** also contributed substantially, reinforcing the impact of diet on the classification problem.

Demographic and behavioral factors, such as **Gender and Family History of Overweight**, showed moderate influence, indicating potential genetic and societal effects.

Less important features, such as **transportation methods (MTRANS), smoking habits (SMOKE_yes), and certain meal frequency variables**, had minimal impact on the model's performance.

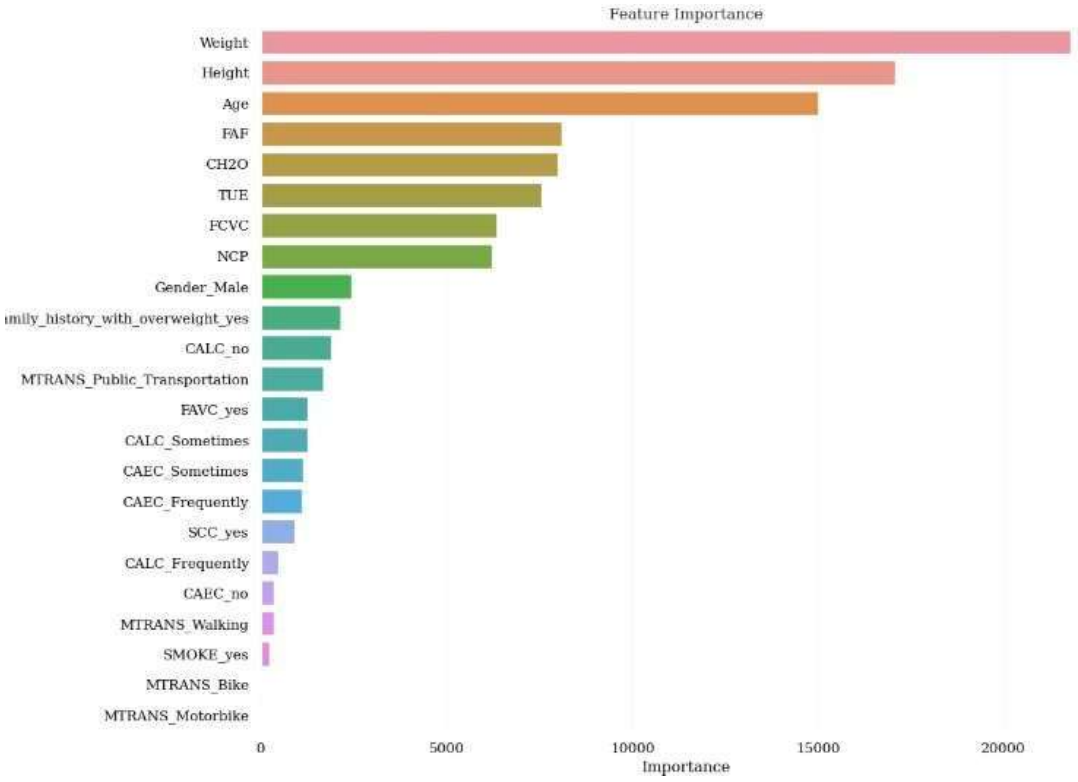

Figure 10: Important features Selection

The performance of five varying machine learning algorithms—Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and Gradient Boosting (LightGBM)—was measured on Accuracy, Precision, Recall, and F1-Score. The findings depict the efficiency of various models for classification tasks.

- Logistic Regression had a moderate predictive capability with an accuracy rate of 0.77. Its 0.857 recall is indicative of its ability to pick up many positive cases, even though precision (0.7519) reveals some misidentification.
- Support Vector Machine (SVM) was slightly higher in accuracy at 0.79, pointing to enhanced generalization. Its precision (0.8152) and recall (0.869) indicate a finer balance in prediction.
- Decision Tree performed much better than the earlier models, achieving an accuracy of 0.91 with good precision (0.9336) and recall (0.943). Decision trees, however, can be overfitting, and additional validation is needed.
- Random Forest, a stronger ensemble model, improved accuracy further to 0.94, with good predictive performance and a balanced precision (0.8905) and recall (0.892).
- Gradient Boosting (LightGBM) was the top-performing model with a high accuracy of 0.97. Its precision (0.9605) and recall (0.972) are high, indicating its ability to reduce false positives and false negatives while having robust overall classification performance.

Table 2: Result of different models after feature selection

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.77	0.7519	0.857	0.7906
Support Vector Machine (SVM)	0.79	0.8152	0.869	0.8308
Decision Tree	0.91	0.9336	0.943	0.9314
Random Forest	0.94	0.8905	0.892	0.8206
Gradient Boosting (LightGBM)	0.97	0.9605	0.972	0.9606

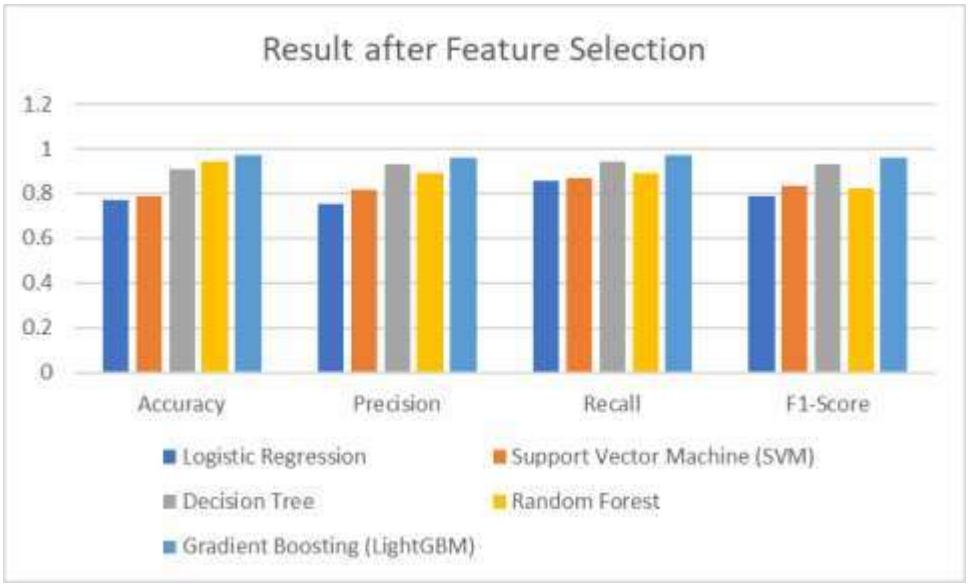

Figure 11: Result of different models after feature selection

4. CONCLUSION AND FUTURE WORK

This research compared different machine learning models for obesity classification based on demographic, lifestyle, and dietary variables, and determined that Weight, Height, and Age were the most significant features. Gradient Boosting (LightGBM) was the best-performing model with 97% accuracy, followed by Random Forest (94%) and Decision Tree (91%), while Logistic Regression (77%) and SVM (79%) were moderately performing. Feature selection had a considerable impact on model accuracy, as it was established that the elimination of irrelevant variables improves predictive performance. The results emphasize the high influence of nutritional patterns, exercise, and heredity on the risk for obesity, highlighting the importance of data-driven intervention. Deep learning and predictive systems in real-time could be included in future studies for enhanced health tracking and prevention of early obesity. Future research can delve into the application of deep learning models, e.g., neural networks, for improving predictive accuracy and identifying intricate relationships among factors related to obesity. Integration with real-time tracking systems based on wearable devices and IoT technology could enable ongoing health monitoring and facilitate early detection of obesity. Larger datasets with varying populations and longitudinal data can further improve model generalization and identify changing trends in obesity. Additional studies can also concentrate on explainable AI (XAI) methods for enhancing model explainability and establishing trust in healthcare. Finally, creating a personalized recommendation system from predictive insights can assist individuals in making sound lifestyle decisions for preventing obesity.

REFERENCES:

[1] An, R., Shen, J., & Xiao, Y. (2022). Applications of artificial intelligence to obesity research: scoping review of methodologies. Journal of Medical Internet Research, 24(12), e40589.
[2] Marshall, T., Champagne-Langabeer, T., Castelli, D., & Hoelscher, D. (2017). Cognitive computing and eScience in health and life science research: artificial intelligence and obesity intervention programs. Health information science and systems, 5, 1-11.
[3] Bouharati, S., Bounechada, M., Djoudi, A., & Harzallah, D. (2012). Prevention of obesity using artificial intelligence techniques. International Journal of Science and Engineering Investigations, 1(9).
[4] Aiosa, G. V., Palesi, M., & Sapuppo, F. (2023). EXplainable AI for decision Support to obesity comorbidities diagnosis. IEEE Access, 11, 107767-107782.
[5] Alghalyini, B. (2023). Applications of artificial intelligence in the management of childhood obesity. Journal of family medicine and primary care, 12(11), 2558-2564.
[6] Thomas, D. M., Knight, R., Gilbert, J. A., Cornelis, M. C., Gantz, M. G., Burdekin, K., ... & Kleinberg, S. (2024). Transforming Big Data into AI-ready data for nutrition and obesity research. Obesity, 32(5), 857-870.
[7] Zarkogianni, K., Chatzidaki, E., Polychronaki, N., Kalafatis, E., Nicolaides, N. C., Voutetakis, A., ... & Nikita, K. (2023). The ENDORSE feasibility study: exploring the use of M-health, artificial intelligence and serious games for the management of childhood obesity. Nutrients, 15(6), 1451.
[8] Allen, B. (2023). An interpretable machine learning model of cross-sectional US county-level obesity prevalence using explainable artificial intelligence. Plos one, 18(10), e0292341.
[9] Rappaport, S. D., & Moskowitz, H. R. (2024). Developing a Mind-Set Framework for Patient-Centered Care on Childhood Obesity, Using AI as a Coach. Children, 1, 3.
[10] Anguita-Ruiz, A., Segura-Delgado, A., Alcalá, R., Aguilera, C. M., & Alcalá-Fdez, J. (2020). eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. PLoS computational biology, 16(4), e1007792.

[11] Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. Current Research in Behavioral Sciences, 2, 100053.

[12] Dugan, T. M., Mukhopadhyay, S., Carroll, A., & Downs, S. (2015). Machine learning techniques for prediction of early childhood obesity. Applied clinical informatics, 6(03), 506-520.

[13] Zheng, Z., & Ruggiero, K. (2017, November). Using machine learning to predict obesity in high school students. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2132-2138). IEEE.

[14] Dirik, M. (2023). Application of machine learning techniques for obesity prediction: a comparative study. Journal of complexity in Health Sciences, 6(2), 16-34.

[15] Rodríguez, E., Rodríguez, E., Nascimento, L., da Silva, A. F., & Marins, F. A. S. (2021, November). Machine learning Techniques to Predict Overweight or Obesity. In IDDM (pp. 190-204).

[16] Pang, X., Forrest, C. B., Lê-Scherban, F., & Masino, A. J. (2021). Prediction of early childhood obesity with machine learning and electronic health record data. International journal of medical informatics, 150, 104454.

[17] Jeon, J., Lee, S., & Oh, C. (2023). Age-specific risk factors for the prediction of obesity using a machine learning approach. Frontiers in Public Health, 10, 998782.

[18] Gerl, M. J., Klose, C., Surma, M. A., Fernandez, C., Melander, O., Männistö, S., ... & Simons, K. (2019). Machine learning of human plasma lipidomes for obesity estimation in a large population cohort. PLoS biology, 17(10), e3000443.

[19] Maulana, A., Afidh, R. P. F., Maulydia, N. B., Idroes, G. M., & Rahimah, S. (2024). Predicting obesity levels with high accuracy: Insights from a catboost machine learning model. Infolitika Journal of Data Science, 2(1), 17-27.

[20] Peng, B., Wu, J., Liu, X., Yin, P., Wang, T., Li, C., ... & Zhang, Y. (2025). Interpretable machine learning for identifying overweight and obesity risk factors of older adults in China. Geriatric Nursing, 61, 580-588.