MACHINE LEARNING APPROACHES FOR PREDICTION OF OBESITY LEVELS BASED ON EATING HABITS

Dr. S. Venkata Achuta Rao[1], K. Mahesh[1*], Ch. Avani Reddy[1], G. Swecha Patel[1], Hemant Pandey[1]

[1]Department of Computer Science and Engineering, Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana, India

ABSTRACT

Obesity is a prevalent global health issue, with multifaceted causes, including genetic, environmental, and lifestyle factors. One significant aspect contributing to obesity is eating habits, making it crucial to understand the relationship between dietary choices and obesity levels. This research explores the application of machine learning (ML) techniques to predict obesity levels based on eating habits. Here, a comprehensive dataset encompassing diverse demographic information, dietary patterns, and obesity levels of individuals is considered. Various machine learning algorithms, including Decision Trees, Support Vector Machines, Random Forests, and Neural Networks, are employed to develop predictive models. Feature selection methods are employed to identify the most influential dietary factors affecting obesity. The proposed approach assesses the model's performance using metrics such as accuracy, precision, recall, and F1-score. Additionally, ML models demonstrate promising predictive capabilities, with certain algorithms outperforming others in accuracy and reliability. Moreover, feature importance analysis identifies specific food groups and consumption patterns strongly associated with obesity, providing valuable insights for targeted interventions and personalized dietary recommendations. This research contributes to the growing field of predictive healthcare analytics, offering a data-driven approach to address obesity-related challenges. The outcomes have implications for public health policies, nutrition education programs, and personalized healthcare initiatives, aiming to mitigate the obesity epidemic and promote healthier lifestyles.

Keywords: Public health, Obesity, Eating habits, Feature selection, Predictive modelling, Machine learning.

1. INTRODUCTION

Obesity has become a global health concern, with rising rates worldwide. It is associated with various health issues such as diabetes, cardiovascular diseases, and certain types of cancer. Understanding and predicting obesity levels based on eating habits is crucial for preventive healthcare and policymaking. Early attempts to understand the link between eating habits and obesity date back to the mid-20th century. Initially, research focused on simple correlations. As technology advanced, data collection methods improved, leading to more nuanced studies. With the rise of computers and data science, ML approaches were introduced to analyze complex relationships between various factors, including eating habits and obesity. ML offers powerful tools to analyze vast datasets and extract meaningful patterns. In predicting obesity levels based on eating habits, ML algorithms process various features like food types, portion sizes, meal timings, and physical activity levels to create predictive models. The need for predicting obesity levels based on eating habits is multifaceted. It aids public health officials in designing effective interventions and policies. Additionally, it helps individuals make informed decisions about their lifestyles, leading to healthier choices and reduced healthcare costs.

Problem Statement

Obesity poses a widespread global health concern, influenced by a complex interplay of genetic, environmental, and lifestyle factors. Among these, eating habits play a pivotal role, underscoring the need to unravel the intricate relationship between dietary choices and obesity levels. This study employs advanced machine learning (ML) techniques to predict obesity levels based on a comprehensive dataset encompassing diverse demographic information, dietary patterns, and individual obesity status.

2. LITERATURE SURVEY

S. Maria [1] proposed that approximately about two billion peoples are affected by obesity that has drawn significant attention on social media. As the sedentary lifestyle which includes consumption of junk foods, no physical activities,spending more on screen,etc are one of the causes of obesity.Obesity generally refers to that a person's body possessing an excessive amount of fat.There is a huge increase in obesity cases which resulting cardiac problems,stroke,insomnia, breathing problems,etc.Type-2 diabetes has been detected in the patients suffering from obesity recently. The studies showing that there are lot of young individuals and children's who has been suffering from overweight and obesity issues in Bangladesh. Here, a strategy for predicting the risk of obesity is proposed that makes use of various machine learning methods. The dataset Obesity and Lifestyle taken from Kaggle site which is collection of different data based on the eating habits and physical conditions,such as height, weight,calorie intake,physical activities are just a few of the 17 different categories in the dataset that reflect the elements that cause obesity. Several machine learning methods include Gradient Boosting Classifier, Adaptive Boosting (ADA boosting), K-nearest Neighbor (K-NN), Support Vector Machine (SVM), Random Forest, and Decision Tree.

T. Cui [2] proposed in recent decades, there has been increasing concern about obesity in adolescents and adults. Obesity can cause many physical health problems and affect people's quality of life. So people are starting to look at the factors that lead to obesity and predict the emergence of obesity. This research presents an estimation of obesity levels based on eating habits, physical condition, and other factors, using a dataset found on UCI Machine Learning Repository. This dataset contains 17 attributes and 2111 records. The labels of this dataset are classified as Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. In this research, three major methods are chosen for prediction: Decision Trees, Logistic Regression, and K Nearest Neighbor. Finally, the result obtained by Decision Trees has the best accuracy.

N. P. Sable [3] encountered More than 2.1 billion people worldwide are shuddering from overweightness or obesity, which represents approximately 30% of the world's population. Obesity is a serious global health problem. By 2030, 41% of people will likely be overweight or obese, if the current trend continues. People who show indications of weight increase or obesity run the danger of contracting life-threatening conditions including type 2 diabetes, respiratory issues, heart disease, and stroke. Some intervention strategies, like regular exercise and a balanced diet, might be essential to preserving a healthy lifestyle. Thus, it is crucial to identify obesity as soon as feasible. We have collected data from sources like schools and colleges within our organization to create our dataset. A vast range of ages is considered and the BMI value is examined in order to determine the level of obesity. The dataset of people with normal BMI and those at risk has an inherent imbalance. The outcomes are collected and showcased via a website which also includes various preventive measures and calculators. The outcomes are promising, and clock an accuracy of about 90%

Singh, B [4] observed individuals developing signs of weight gain or obesity are also at a risk of developing serious illnesses such as type 2 diabetes, respiratory problems, heart disease and stroke. Some intervention measures such as physical activity and healthy eating can be a fundamental component to maintain a healthy lifestyle. Therefore, it is absolutely essential to detect childhood obesity as early as possible. This paper utilises the vast amount of data available via UK's millennium cohort study in order to construct a machine learning driven model to predict young people at the risk of becoming overweight or obese. The childhood BMI values from the ages 3, 5, 7 and 11 are used to predict adolescents of age 14 at the risk of becoming overweight or obese. There is an inherent imbalance in the dataset of individuals with normal BMI and the ones at risk. The results obtained are encouraging and a prediction accuracy of over 90% for the target class has been achieved. Various issues relating to data preprocessing and prediction accuracy are addressed and discussed.

Cheng [5]. used 11 classification algorithms (logistic regression, radial basis function (RBF), naïve Bayes, classification via regression (CVR), local k-nearest neighbors (k-NN), a decision table, random subspace, random tree, a multi-objective evolutionary fuzzy classifier, and a multilayer perceptron) to predict obesity in adults and achieved a highest overall accuracy of 70% with a random subspace algorithm.

Cervantes. [6]. developed decision tree (DT), k-means, and support vector machine (SVM)-based data mining techniques to identify obesity levels among young adults between 18 and 25 years of age so that interventions could be undertaken to maintain a healthier lifestyle in the future. Gupta [7]. developed a deep learning model (long short-term memory (LSTM)), which predicted obesity between 3 and 20 years of age with 80% accuracy using unaugmented electronic health record (EHR) data from 1 to 3 years prior. Marcos-Pasero [8] used random forest (RF) and gradient boosting to predict the BMI from 190 multidomain variables (data collected from 221 children aged 6 to 9 years) and determined the relative importance of the predictors. Zare [9]. used kindergarten-level BMI information, demographic, socioeconomic information such as family income, poverty level, race, ethnic compositing, housing, parent education, and family structure to predict obesity at the fourth grade and achieved an accuracy of about 87% by using logistic regression and an artificial neural network. Zare [10]. used kindergarten-level BMI information, demographic, socioeconomic information such as family income, poverty level, race, ethnic compositing, housing, parent education, and family structure to predict obesity at the fourth grade and achieved an accuracy of about 87% by using logistic regression and an artificial neural network. Fu, et. al [11]. developed an ML-based framework to predict childhood obesity by using health examination, lifestyle and dietary habits, and anthropometric measurement-related data.

3. PROPOSED METHODOLOGY

Dataset Collection and Characteristics: A dataset incorporating a broad spectrum of information is collated, capturing demographic details, dietary preferences, and obesity levels of individuals. This diverse dataset forms the foundation for training and evaluating machine learning models.

Data Preprocessing: To ensure the reliability of the dataset, a meticulous data preprocessing phase is undertaken. This involves handling missing values, standardizing data formats, and addressing outliers. Cleaning the dataset lays the groundwork for accurate and meaningful model training.

Data Splitting: The dataset is then partitioned into training and testing sets. The training set is utilized to teach the machine learning models, while the testing set evaluates the model's predictive performance on unseen data, simulating real-world scenarios.

Machine Learning Algorithms: Various machine learning algorithms, including Decision Trees, Support Vector Machines, Random Forests, and Neural Networks, are employed. Each algorithm is trained on the dataset, learning patterns and relationships between dietary factors and obesity levels.

Feature Selection Methods: Feature selection methods are applied to identify the most influential dietary factors affecting obesity. This step is crucial for refining the model and focusing on the most relevant features, improving both accuracy and interpretability.

Performance Evaluation: The performance of the machine learning models is rigorously assessed using key metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the model's effectiveness in predicting obesity levels based on eating habits.

Prediction from Test Data: The trained models are then applied to the test dataset to predict obesity levels based on individuals' eating habits. This step is crucial for gauging the real-world applicability and reliability of the developed models.

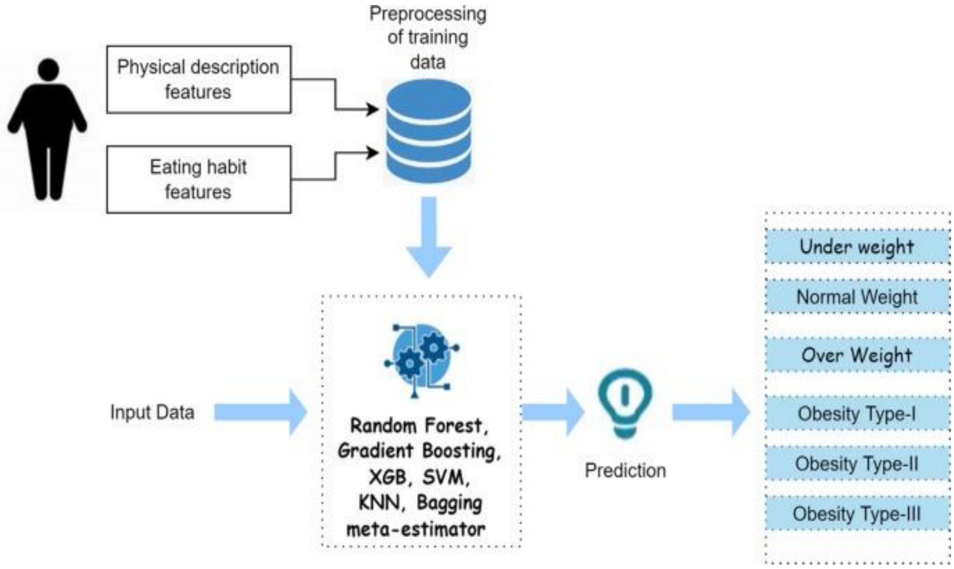

Preprocessing of training data
Physical description features
Eating habit features
Input Data
Random Forest, Gradient Boosting, XGB, SVM, KNN, Bagging meta-estimator
Prediction
Under weight
Normal Weight
Over Weight
Obesity Type-I
Obesity Type-II
Obesity Type-III

Figure 1: Proposed system architecture.

Extremely Randomized Trees

Extremely Randomized Trees, also known as Extra Trees, construct multiple trees like RF algorithms during training time over the entire dataset. During training, the ET will construct trees over every observation in the dataset but with different subsets of features.

It is important to note that although bootstrapping is not implemented in ET's original structure, we can add it in some implementations. Furthermore, when constructing each decision tree, the ET algorithm splits nodes randomly.

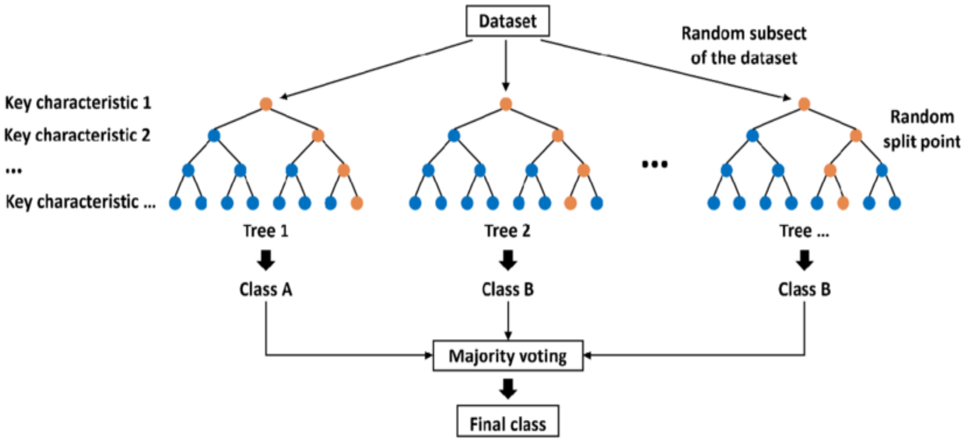
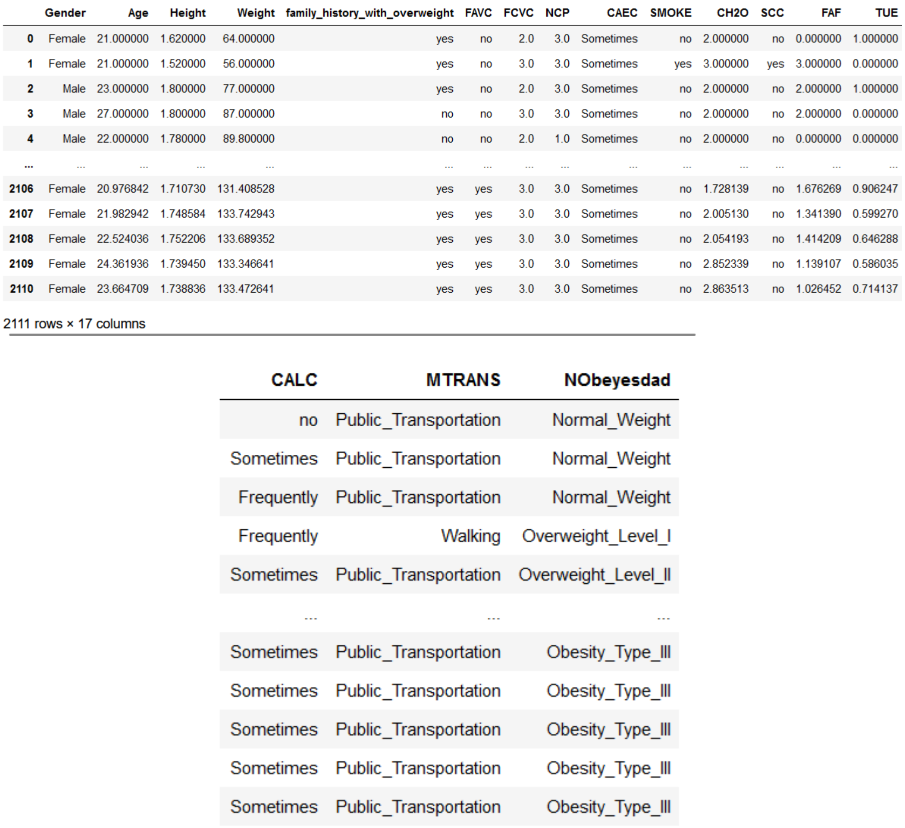


Figure 2: diagram of Extremely Randomized Tree Classifier.

4. RESULTS AND DISCUSSION

Figure 3 provides a visual representation or illustration of a sample dataset used for predicting obesity levels based on eating habits. It includes various data points or examples from the dataset, each characterized by different features such as gender, age, height, weight, and other relevant factors. The goal is to visualize how the data is structured and what kind of information it contains. It helps viewers get an overview of the diversity and distribution of data points in the dataset.



	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE
0	Female	21.000000	1.620000	64.000000	yes	no	2.0	3.0	Sometimes	no	2.000000	no	0.000000	1.000000
1	Female	21.000000	1.520000	56.000000	yes	no	3.0	3.0	Sometimes	yes	3.000000	yes	3.000000	0.000000
2	Male	23.000000	1.800000	77.000000	yes	no	2.0	3.0	Sometimes	no	2.000000	no	2.000000	1.000000
3	Male	27.000000	1.800000	87.000000	no	no	3.0	3.0	Sometimes	no	2.000000	no	2.000000	0.000000
4	Male	22.000000	1.780000	89.800000	no	no	2.0	1.0	Sometimes	no	2.000000	no	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2106	Female	20.976842	1.710730	131.408528	yes	yes	3.0	3.0	Sometimes	no	1.728139	no	1.676269	0.906247
2107	Female	21.982942	1.748584	133.742943	yes	yes	3.0	3.0	Sometimes	no	2.005130	no	1.341390	0.599270
2108	Female	22.524036	1.752206	133.689352	yes	yes	3.0	3.0	Sometimes	no	2.054193	no	1.414209	0.646288
2109	Female	24.361936	1.739450	133.346641	yes	yes	3.0	3.0	Sometimes	no	2.852339	no	1.139107	0.586035
2110	Female	23.664709	1.738836	133.472641	yes	yes	3.0	3.0	Sometimes	no	2.863513	no	1.026452	0.714137

2111 rows × 17 columns

CALC	MTRANS	NObeyesdad
no	Public_Transportation	Normal_Weight
Sometimes	Public_Transportation	Normal_Weight
Frequently	Public_Transportation	Normal_Weight
Frequently	Walking	Overweight_Level_I
Sometimes	Public_Transportation	Overweight_Level_II
...	...	...
Sometimes	Public_Transportation	Obesity_Type_III
Sometimes	Public_Transportation	Obesity_Type_III
Sometimes	Public_Transportation	Obesity_Type_III
Sometimes	Public_Transportation	Obesity_Type_III
Sometimes	Public_Transportation	Obesity_Type_III

Figure 3: Illustration of sample dataset used for obesity level prediction.

Figure 4 is a graphical representation, specifically a count plot, of the distribution of the target label in the dataset. The target label in this context is probably the "NObeyesdad" column, which represents the obesity level. The count plot visualizes how many instances or data points belong to each category of obesity level. Each category on the x-axis (e.g., Normal, Overweight, etc.) will have a corresponding bar indicating the count or frequency of occurrences in the dataset. This type of plot is useful for understanding the balance or imbalance in the distribution of different classes, which is crucial for classification tasks in machine learning.

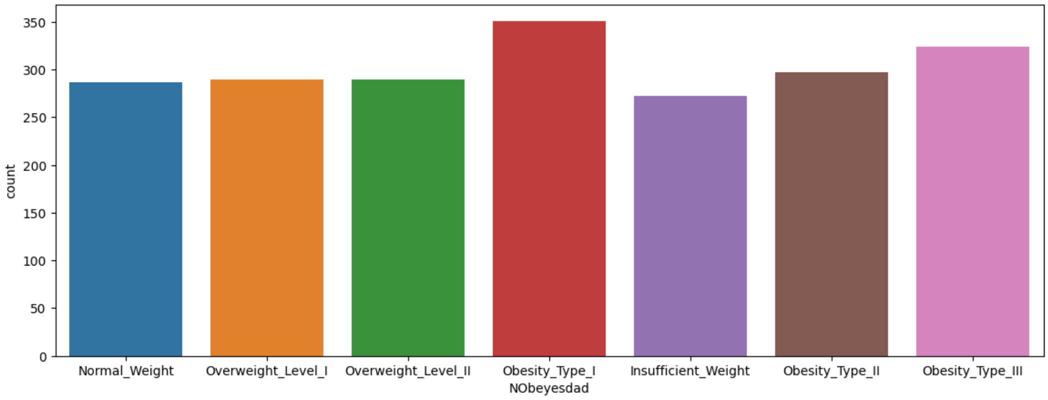


Figure 4: Displaying the count plot for target label.

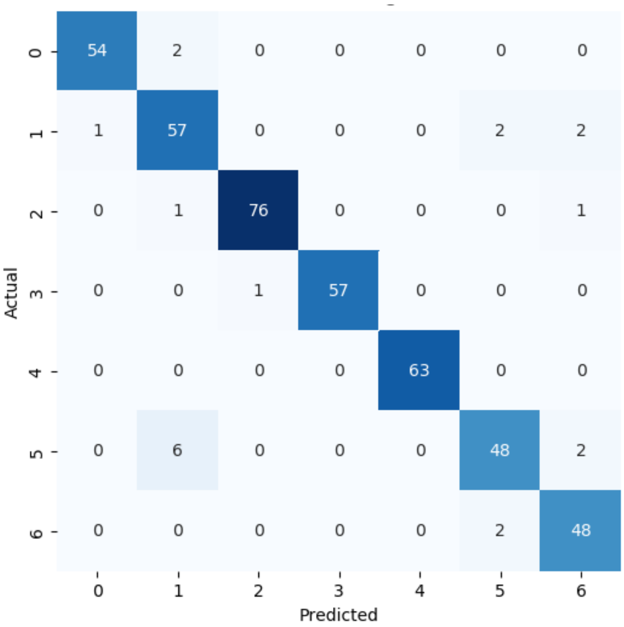


Figure 5: Displays the confusion matrix of Extra Trees Classifier model.

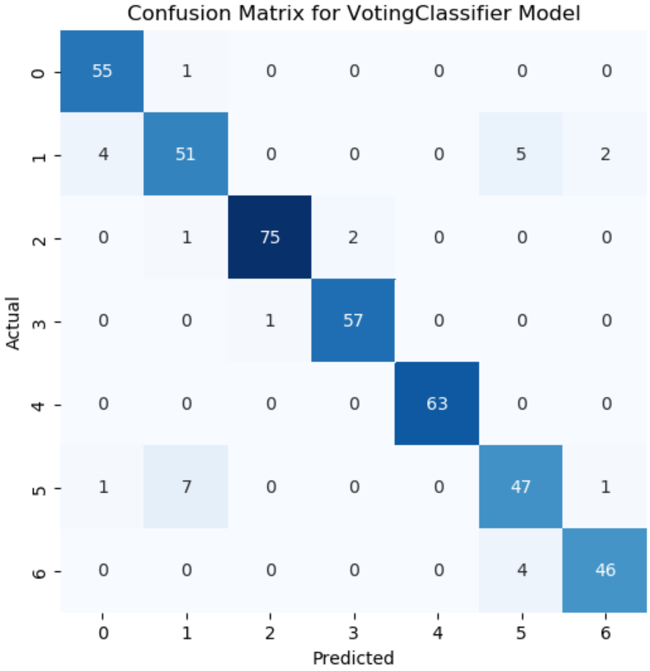


Figure 6: Displays the confusion matrix of voting classifier model.

Table 1: Performance comparison of existing and proposed ML models.

Model	Accuracy (%)	Precision	Recall	F1-score
Voting Classifier	0.93	0.93	0.93	0.93
ETC model	0.95	0.95	0.95	0.95

For the Voting Classifier model:

- The Accuracy is 0.93, indicating the accuracy between the actual and predicted values
- The Precision is 0.93, suggesting that, on average Precision between the actual and predicted values.
- The Recall is 0.93, suggesting that, on average Recall between the actual and predicted values.
- The F1-score is 0.93, representing the average F1-score between the actual and predicted values.

For the Extra Tree Classifier model:

- The Accuracy is 0.95, indicating the accuracy between the actual and predicted values.
- The Precision is 0.95, suggesting that, on average Precision between the actual and predicted values.
- The Recall is 0.95, suggesting that, on average Recall between the actual and predicted values.
- The F1-score is 0.95, representing the average F1-score between the actual and predicted values.

5. CONCLUSION

In conclusion, this research has delved into the intricate link between eating habits and obesity, leveraging the power of machine learning (ML) to predict obesity levels. The utilization of diverse demographic information and dietary patterns within a comprehensive dataset has allowed for a nuanced exploration of this complex health issue. Through the application of various ML algorithms such as Decision Trees, Support Vector Machines, Random Forests, and Neural Networks, predictive models have been crafted, revealing the potential of data-driven methodologies in understanding and addressing obesity. The evaluation of model performance using metrics like accuracy, precision, recall, and F1-score has provided a robust assessment of the predictive capabilities of these ML algorithms. The findings indicate promising outcomes, with certain algorithms showcasing superior accuracy and reliability in forecasting obesity levels based on eating habits. Moreover, the incorporation of feature selection methods has unveiled the most influential dietary factors contributing to obesity. This not only enhances the interpretability of the models but also provides actionable insights for targeted interventions. The identification of specific food groups and consumption patterns strongly associated with obesity adds a layer of granularity, paving the way for personalized dietary recommendations and more effective strategies to combat the obesity epidemic.

REFERENCES

[1] S. Maria, R. Sunder and R. S. Kumar, "Obesity Risk Prediction Using Machine Learning Approach," 2023 International Conference on Networking and Communications (ICNWC), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICNWC57852.2023.10127434.

[2] T. Cui, Y. Chen, J. Wang, H. Deng and Y. Huang, "Estimation of Obesity Levels Based on Decision Trees," 2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM), Xi'an, China, 2021, pp. 160-165, doi: 10.1109/ISAIAM53259.2021.00041.

[3] N. P. Sable, R. Bhimanpallewar, R. Mehta, S. Shaikh, A. Indani and S. Jadhav, "A Machine Learning approach for Early Detection and Prevention of Obesity and Overweight," 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 2023, pp. 1-5, doi: 10.1109/I2CT57861.2023.10126346.

[4] Singh, B., Tawfik, H. (2020). Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People. In: Krzhizhanovskaya, V.V., et al. Computational Science – ICCS 2020. ICCS 2020. Lecture Notes in Computer Science(), vol 12140. Springer, Cham.

[5] Cheng, X.; Lin, S.-y.; Liu, J.; Liu, S.; Zhang, J.; Nie, P.; Fuemmeler, B.F.; Wang, Y.; Xue, H. Does physical activity predict obesity—A machine learning and statistical method-based analysis. Int. J. Environ. Res. Public Health 2021, 18, 3966

[6] Cervantes, R.C.; Palacio, U.M. Estimation of obesity levels based on computational intelligence. Inform. Med. Unlocked 2020, 21, 100472.

[7] Gupta, M.; Phan, T.-L.T.; Bunnell, H.T.; Beheshti, R. Obesity Prediction with EHR Data: A deep learning approach with interpretable elements. ACM Trans. Comput. Healthc. (HEALTH) 2022, 3, 1–19.

[8] Marcos-Pasero, H.; Colmenarejo, G.; Aguilar-Aguilar, E.; Ramírez de Molina, A.; Reglero, G.; Loria-Kohen, V. Ranking of a wide multidomain set of predictor variables of children obesity by machine learning variable importance techniques. Sci. Rep. 2021, 11, 1910

[9] Zare, S.; Thomsen, M.R.; Nayga Jr, R.M.; Goudie, A. Use of machine learning to determine the information value of a BMI screening program. Am. J. Prev. Med. 2021, 60, 425–433

[10] Fu, Y.; Gou, W.; Hu, W.; Mao, Y.; Tian, Y.; Liang, X.; Guan, Y.; Huang, T.; Li, K.; Guo, X. Integration of an interpretable machine learning algorithm to identify early life risk factors of childhood obesity among preterm infants: A prospective birth cohort. BMC Med. 2020, 18, 184

[11] Pang, X.; Forrest, C.B.; Lê-Scherban, F.; Masino, A.J. Prediction of early childhood obesity with machine learning and electronic health record data. Int. J. Med. Inform. 2021, 150, 104454.