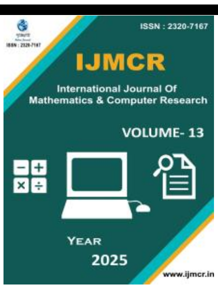
International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 13 Issue 05 May 2025, Page no. – 5144-5150

Index Copernicus ICV: 57.55, Impact Factor: 8.615

DOI: 10.47191/ijmcr/v13i5.01



Prediction of Obesity Among Adults and Adolescents Using a Machine Learning Approach

Md. Hamidul Islam[1], Md. Shohel Rana (PhD)[2]

[1,2]Department of Statistics, Islamic University, Kushtia-7003, Bangladesh

ARTICLE INFO

Published Online:
01 May 2025

Corresponding Author:
Md. Shohel Rana (PhD)

ABSTRACT

Obesity has become one of the most significant public health issues of the 21st century. Obesity is a chronic, complex disease characterized by excessive fat accumulation that impairs health. It also has social and psychological dimensions, impacting individuals across all age groups and socioeconomic strata. It is associated with a range of risk factors, including diabetes, depression, and cancer. It presents a significant challenge to both developed and developing countries worldwide. This study aims to explore the factors contributing to obesity and develop a predictive model to identify individuals at risk. Using a secondary dataset, three machine learning algorithms were implemented for both interpretation and prediction. Logistic regression was used to identify significant associations between obesity and key factors such as family history of overweight and high caloric food intake. Among the predictive models, the random forest classifier demonstrated superior performance. The model was evaluated using metrics such as accuracy, specificity, and the Receiver Operating Characteristic (ROC) curve. The results indicated that the random forest classifier was the most effective model for predicting obesity, achieving the highest accuracy and ROC values. In conclusion, the findings of this study suggest that machine learning models, particularly the random forest classifier, can be effectively used to identify at-risk individuals and may offer valuable insights for the healthcare sector. The integration of such models could improve targeted interventions and support public health initiatives aimed at mitigating the obesity epidemic.

KEYWORDS: Obesity, Machine Learning, Random Forest, Logistic Regression, Obesity Epidemic

I. INTRODUCTION

Obesity has become one of the most significant public health issues of the 21st century. WHO says it is a chronic complex disease and cases of type 2 diabetes and heart disease. Sometimes it causes of certain cancer also. Obesity influences the quality of living, such as sleeping or moving and in 2022, one in 8 people in the world were living with obesity Worldwide obesity has nearly tripled since 1975. In 2016, more over 18 years aged adult about 39% were overweight and 13% were obese. In just 40 years, the number of school-age children and adolescents with obesity has risen more than tenfold, from 11 million to 124 million [1]. This trend is not confined to high-income countries; it is increasingly evident in low- and middle-income countries, contributing to a dual burden of malnutrition and obesity. If it keeps as similar manner, [2] explains one third of world adult population will be overweight and obese will be about 1 billion by 2025. For a long period of time if energy intake increases energy expenditure, obesity raised and increased body fat [3]. The causes of obesity are complex and multifaceted, involving a mix of genetic, behavioral, environmental, and metabolic factors. Major contributors include unhealthy eating habits, lack of physical activity, socioeconomic factors, and specific medical conditions. In 2019, higher-than-optimal BMI (Body Mass Index) caused an estimated 5 million deaths from non-communicable diseases (NCDs) such as cardiovascular diseases, diabetes, cancers, neurological disorders, chronic respiratory diseases, and digestive disorders [4]. The overall challenge of both underweight and obesity has risen in the majority of nations, prompted by a rise in obesity, whereas underweight and thinness continue to be widespread in South Asia and certain regions of Africa explained by [5]. Predicting obesity using machine learning is an emerging field of research that aims to utilize the extensive health data available to identify individuals at risk of becoming obese. By examining patterns and correlations within the data, machine learning models can offer insights into the factors leading to obesity and predict future cases with a level of accuracy that traditional statistical methods may not match. In this study we emphasized more on lifestyle factors without denying the genetics like family history of overweight. Besides the patterns found in this study meets our intuition which makes it more acceptable and more comprehensive.

II. LITERATURE REVIEW

Researchers from multiple studies have focused on predicting obesity using various machine learning algorithms. [6] Investigates the impact of different factor quantities on predicting obesity status using five machine learning approaches: Decision Tree, Support Vector Machine, Extreme Gradient Boosting, Random Forest, and Extremely Randomized Trees models. The study focuses on 2111 observations related to eating habits and behavior patterns of obesity, reflecting the current interest in addressing obesity as a chronic disease affecting daily life. The results highlight that the XGBoost model performs the best with an accuracy of 97.16% under 14 factors, indicating the effectiveness of this approach in predicting obesity status. [7] Explores the use of machine learning to predict obesity based on various lifestyle factors. Initially, an exploratory data analysis was conducted to identify patterns and relationships between variables and obesity. A simple random forest model was then used to establish a baseline for variable importance. [8] Focused on supervised learning methods, including logistic regression, random forest, gradient boosting, and XGBoost, prioritizing predictability over interpretability, except for logistic regression, which provided insights into how predictors impact obesity. Logistic regression revealed that individuals using automobiles or public transport are more likely to be obese compared to those walking or biking. The XGBoost model achieved an 83.9% accuracy in predicting obesity. The study aimed to exceed the 50% accuracy benchmark, considering that obesity is influenced equally by genetics and environment, and surpassed this goal by 34%.

[9] Discussed in depth about the causes of obesity and criticized about BMI as an indicator of obesity and proposed a more practical definition of obesity as - Obesity is characterized by an excessive buildup of body fat that negatively impacts health. They also discussed about the link of type 2 diabetes with insulin resistance and β-cell dysfunction, with obesity being a major risk factor due to excess body fat, especially in the abdominal region. Obesity is strongly influenced by energy imbalance, where caloric intake exceeds energy expenditure, and is regulated by complex systems involving the hypothalamus and genetic predisposition. Leptin, a key hormone in regulating appetite and energy, plays a crucial role in this process, but its effectiveness in treating obesity is limited. Genetic and environmental factors, such as diet and reduced physical activity, also contribute to obesity. The increasing prevalence of obesity worldwide is largely driven by social and environmental changes. Addressing obesity is vital for preventing type 2 diabetes, requiring a multifactorial approach, including lifestyle changes, targeted interventions, and possibly medical treatments.

[10] Did a work titled "A machine learning approach for obesity risk prediction” applied 9 different machine learning model and found logistic regression most efficient. [11] Explains obesity, in turn, is caused by the accumulation of excess fat. Obesity is influenced by a variety of factors, including age, weight, height, and body mass index. While there are various methods to calculate obesity, these approaches are not universally applicable in all situations (like for a pregnant woman or an elderly man) and still yield precise outcomes.

This paper focuses on predicting obesity in Bangladesh using various machine learning techniques. Data was collected to predict obesity risk through nine different classifiers, evaluated based on six performance metrics. Logistic regression achieved the highest accuracy of 97.09%. Future plans involve expanding the dataset to include a wider range of low, medium, and high-obesity cases for more comprehensive analysis.

III. METHODOLOGY

A. *Introduction*

We have used a secondary data from the UCI Machine Learning Repository. After the necessary pre-processing step we applied three ML algorithm Logistic regression, Random Forest and Support Vector Machine.

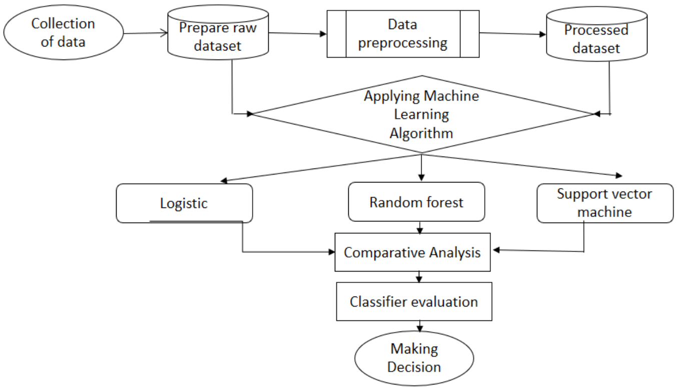


Fig 1. Flowchart of the work process

B. *Logistic Regression*

Logistic regression is a supervised machine learning technique used for classification tasks, specifically to predict binary or categorical outcomes. Unlike linear regression, which is used for continuous variables, logistic regression is suited for predicting the likelihood of a binary event. In this research, logistic regression is employed to predict whether or not a person is obese. If the objective were to predict a continuous variable like BMI, multiple linear regression would be more appropriate. Logistic regression predicts the probability of an outcome, which is then used to classify an observation. For example, the model might predict a 55% probability that an individual is obese. Depending on the chosen threshold (commonly 50%), this probability is used to classify the individual as obese or not.

The hypothesis for logistic regression is defined by the formula:

$$h_\theta(x) = g(\theta^T x)$$

Where the function g is the sigmoid function:

$$g(z) = \frac{1}{1 + e^{-z}}$$

The sigmoid function calculates the probability of the outcome, resulting in values ranging from 0 to 1 shown in Figure 2.

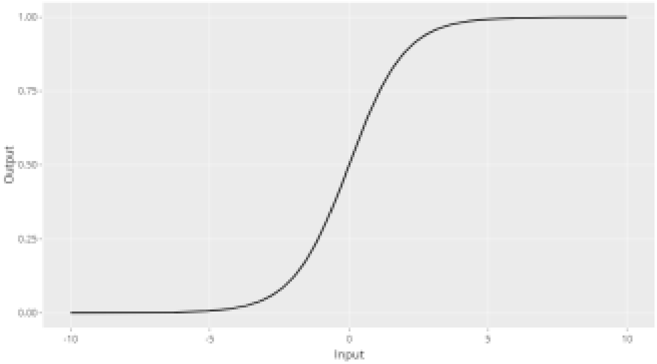


Fig 2. Visual Sigmoid Function Plotted

In developing the logistic regression model for this study, we will start with a baseline model that includes all eligible variables. Following the baseline model creation, variable selection will be conducted using insights from a baseline random forest model. The mean decrease in Gini and mean decrease in accuracy from the random forest model will rank variables based on their importance. The most robust model, determined through this iterative process, will then be compared against other top-performing models to select the best one.

C. *Random Forest*

Random Forest is an ensemble method that can handle both regression and classification tasks by using multiple decision trees and a technique called Bootstrap and Aggregation, or "bagging." The core concept is to combine the outputs of several decision trees to determine the final prediction, rather than depending on individual trees. In Random Forest, multiple decision trees serve as base models. The algorithm randomly samples rows and features from the dataset to create different subsets for each model, a process known as Bootstrap. Random Forest is widely used in industries due to its strong performance and the fact that it does not require any assumptions about data distribution. It is also well-suited for high-dimensional datasets, as is the case in our study.

Before delving into the Random Forest algorithm, it’s important to briefly explain decision trees. There are two types: classification trees and regression trees. Since our dataset involves continuous numeric responses, the regression tree is the appropriate choice. As depicted in the figure, the regression tree processes each node to determine the branch an observation will follow, and when it reaches the bottom of the tree, the group average is used for the final prediction.

Now, we can consider random forest as a combination of multiple independent decision trees, which showed in Figure 3. Each individual tree will go through the nodes independently with the same weight and come up with its own predictions. For the regression, the average of predictions from all decision trees will be used as our final predictions. Also, there are some unique characteristics of random forest. Instead of using all observations, we can randomly choose n bootstrap samples from the whole observations, and for each bootstrap sample, only m out of the total number of features (specified as mtry in R) will be used in each node, which helps to prevent from over fitting [3].

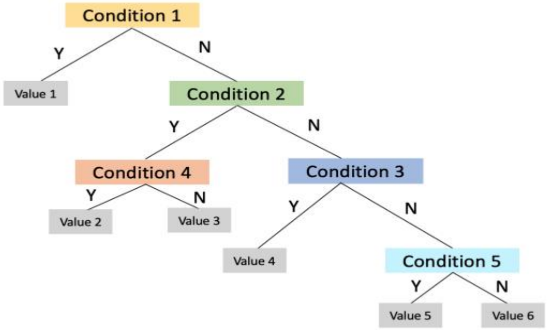


Fig 3. Decision Tree

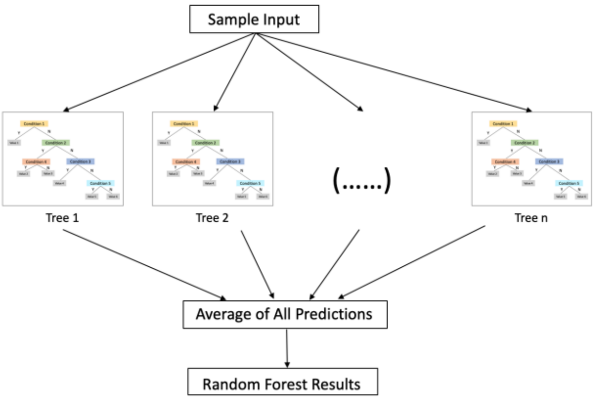


Fig 4. Aggregation of Decision Tree

Random Forest can be viewed as a combination of multiple independent decision trees, as illustrated in Figure 4. Each tree processes the nodes separately, assigning equal weight, and generates its own prediction. In regression tasks, the final prediction is the average of all individual tree predictions. One key feature of Random Forest is its ability to randomly select n bootstrap samples from the entire dataset, rather than using all observations. Additionally, at each node, only m features (defined by mtry in R) are randomly chosen from the total feature set, which helps reduce the risk of over fitting.

D. Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for both classification and regression tasks, particularly well-suited for classification. Unlike logistic regression and random forest, SVM focuses on finding the optimal hyper plane that maximally separates classes in the feature space. This method is highly effective for high-dimensional spaces and cases where the number of dimensions exceeds the number of samples. SVM operates by finding the hyper plane that has the largest margin, meaning the greatest distance between the hyper plane and the nearest data points from each class, known as support vectors. The objective is to maximize this margin to improve the model's generalizability and reduce the risk of misclassification.

The methodology for using SVM in this study involves several steps, starting from data preprocessing to hyper parameter tuning and model evaluation. Data preprocessing involves cleaning the dataset to handle missing values and outliers, ensuring the quality of the input data. The dataset is split into training and testing sets, typically with an 80/20 ratio. The SVM model is then trained on the training data to find the optimal hyper plane. The choice of kernel function is critical in SVM as it defines the transformation applied to the data to find the hyper plane in higher dimensions. Frequently used kernel functions consist of linear, polynomial, radial basis function (RBF), and sigmoid. Initially, a linear kernel is used to create a baseline model. If the linear kernel does not perform well, more complex kernels like RBF are explored.

Hyper parameter tuning is crucial for improving the performance of the SVM model. Key hyper parameters include the regularization parameter (C), which controls the trade-off between achieving a low error on the training data and minimizing the margin, and kernel parameters, such as gamma for the RBF kernel, which defines the influence of a single training example. Grid search and cross-validation techniques are employed to find the optimal combination of these hyper parameters. The performance of the SVM model is evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy measures the proportion of correctly classified instances, precision the proportion of true positive predictions among all positive predictions, recall the proportion of true positive predictions among all actual positives, F1-score the harmonic mean of precision and recall, and ROC-AUC the area under the Receiver Operating Characteristic curve, providing a measure of reparability.

E. Data Collection

We have used a secondary data for this study. The data consist of the estimation of obesity levels in people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition, data was collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 17 attributes and 2111 records. Except Age, Height and Weight there are no other numeric variable. The attributes related with eating habits are: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), and Time using technology devices (TUE), Transportation used (MTRANS). There are 7 different values for the variable (NObeyesdad). The Obesity values are: Underweight Less than 18.5; Normal 18.5 to 24.9; Overweight 25.0 to 29.9; Obesity I: 30.0 to 34.9; Obesity II: 35.0 to 39.9; Obesity III: Higher than 40. The data contains both numerical and continuous data, so it can be used for analysis based on algorithms of classification. The Variable types and their descriptions are shown in table 1.

Table 1: Variables and Description

Variables	Description
Age	Age of individuals
Height	Height of individuals
Gender	Male or Female
family_history_ with_overweight	Yes or No value
FAVC	Frequent consumption of high caloric food
FCVC	Frequency of consumption of vegetables
NCP	Number of main meals
CAEC	Consumption of food between meals
SMOKE	Smoker or not
CH20	Consumption of water daily
SCC	Calories consumption monitoring
FAF	Physical activity frequency
TUE	Time using technology devices
CALC	Consumption of alcohol
MTRANS	Transportation used
NObeyesdad	7 types of obesity values

F. Computational Software

Machine learning algorithm requires special packages or libraries. All the programs such as "R version 4.3.2 (2023-

10-31 ucrt), R Studio (version: ‘2024.4.1.748’) and Microsoft office 2016” are used to generate the results, visualizations and result in this study.

IV. RESULTS AND DISCUSSIONS

A. Exploratory Data Analysis

The actual dataset has 7 different values for the variable ‘NObeyesdad’. We transformed this into just two which are obese and not obese and the new variable is ‘obesity’.

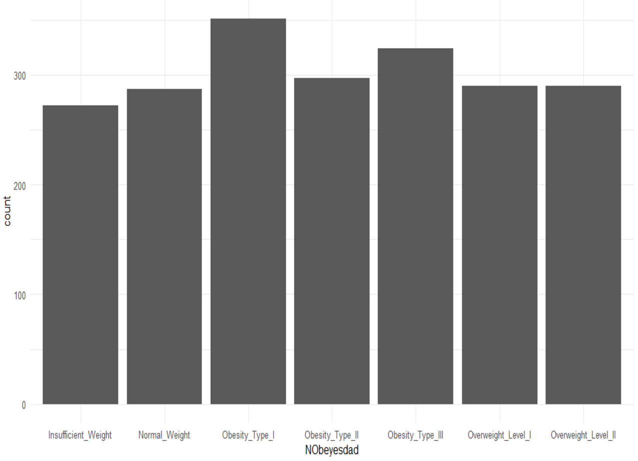


Fig 5. Values of NObeyesdad Variable

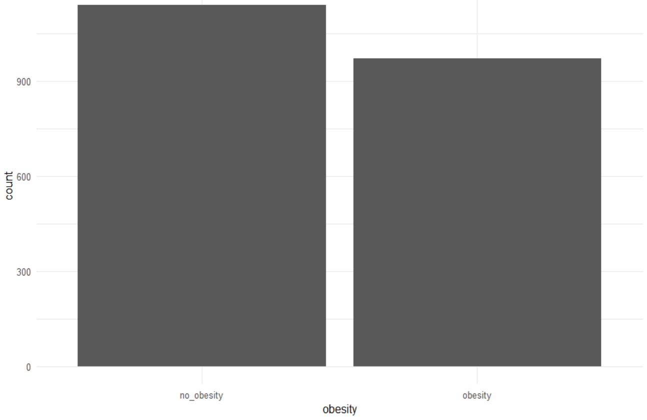


Fig 6. Converted Obesity Variable

At first, we ran a logistic regression and plotted the significant variable against the BMI which is a marker for BMI [2]. It made the interpretation much easier. Let’s see the histogram of BMI.

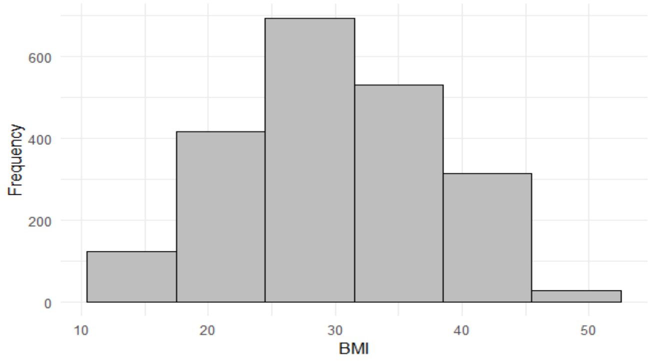


Fig 7. Histogram of BMI

We can see the histogram in Figure 7 is roughly normal. At first we ran a logistic regression model and found age, family history of overweight, frequently consumption of high caloric food (FAVC), frequent consumption of vegetables (FCVC), daily calorie monitoring (SCC) and Physical activity (FAF) to be significant. We plotted these variables against BMI to confirm that is it random or evident.

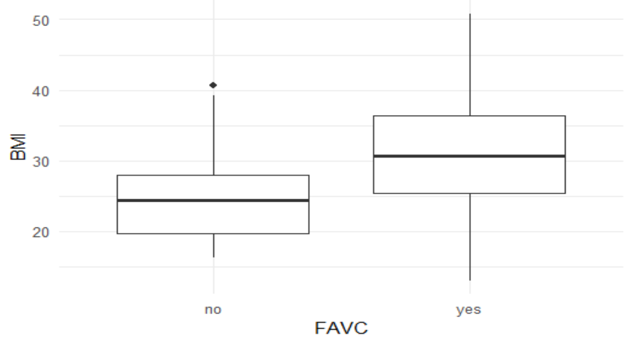


Fig 8. Boxplot of FAVC vs BMI

Figure 8 suggests that median BMI is higher for the individuals who frequently consumes high caloric food (FAVC).

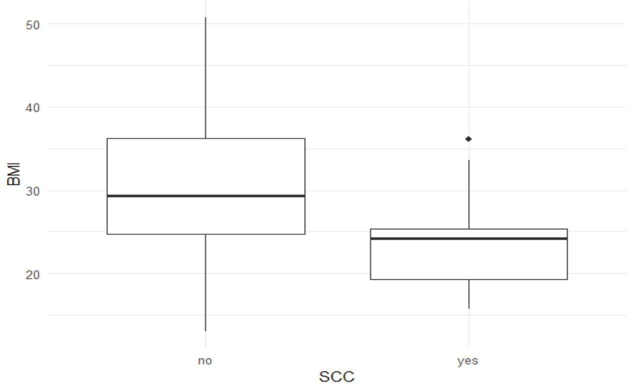


Fig 9. Boxplot of SCC

Figure 9 suggests that median BMI is higher among individuals who monitor their daily calorie intake and expenditure.

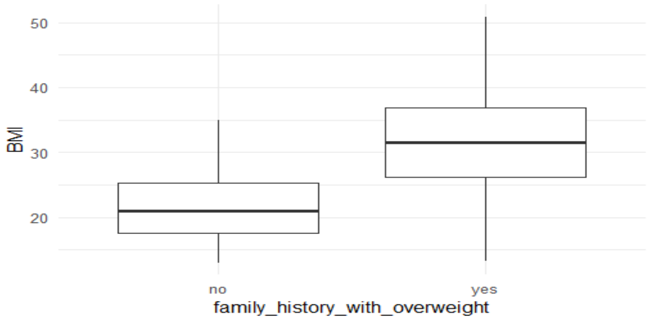


Fig 10. Boxplot of Family History vs BMI

In this study, we compared the performance of three machine learning algorithms—Logistic Regression, Random

Forest, and Support Vector Machine (SVM)—for predicting obesity. After training and evaluating each model, Random Forest emerged as the best-performing classifier with a remarkable accuracy score of 95%. Below are the detailed results for each model.

B. *Logistic Regression*

The logistic regression model, while effective in handling binary classification problems, achieved an accuracy of 76% with an Area under Curve (AUC) value of 0.8455.

C. *Support Vector Machine (SVM):*

The SVM model, which is known for its robustness in high-dimensional spaces, achieved an accuracy of 86%, which is 10% more than logistic model with an AUC value of 0.9413.

D. *Random Forest:*

The Random Forest model outperformed both Logistic Regression and SVM, achieving a high accuracy score of 92%. This ensemble learning method, which builds multiple decision trees and aggregates their results, demonstrated superior performance in terms of both accuracy and robustness. The use of bootstrapping and feature randomization in Random Forest helped in reducing variance and avoiding over fitting. The AUC value is also the highest for this model which is 0.9709.

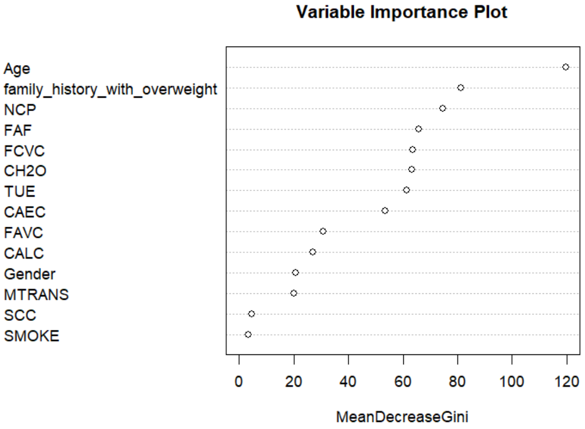


Fig 11. Variable Importance Plot

This is the variable importance plot of our final random forest model. This is calculated by mean decrease gini score which is a measure of how much each variable contributes to the homogeneity of the nodes and leaves in the resulting model. The higher value means higher importance of the variable. The most important variable is age and least important variable is smoke.

E. *Model Comparison*

Table 2: Model Comparison

Model	Accuracy	Specificity	Sensitivity	AUC
Logistic	0.7626	0.8316	0.7038	0.8455
Random Forest	0.9272	0.9106	0.9413	0.9709
Support Vector Machine	0.8655	0.8762	0.8563	0.9413

We have been plotted the receiver operating curves (ROC) of these three model that is shown in Figure 12 and it is evident that random forest is the best classifier for obesity prediction in case of our data.

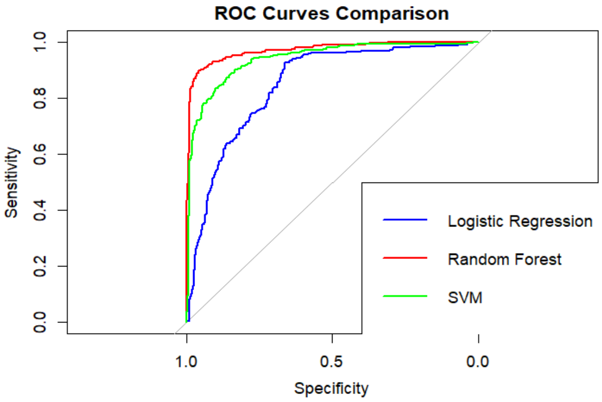


Fig 12. ROC curve

V. CONCLUSION AND FUTURE WORKS

In conclusion, the Random Forest model emerged as the best classifier for predicting obesity in this study, achieving a 95% accuracy score. Its balance of accuracy, robustness, and interpretability makes it a valuable tool for healthcare professionals aiming to identify individuals at risk of obesity and develop targeted intervention strategies. Future work could explore the integration of additional data sources and the application of other advanced machine learning techniques to further enhance predictive performance and practical applicability. Since the demographics are different based on country, so, if we can collect data from Bangladesh, it will be more effective for us to effectively take intervention strategies and policy making. Future recommendations for enhancing obesity prediction models include integrating diverse data sources, facilitating translation into clinical practice, and prioritizing ethical considerations. These efforts aim to improve prediction accuracy, model interpretability, and real-world applicability, ultimately contributing to better obesity management and population health outcomes.

CONFLICT OF INTEREST

Having no conflict of interest declared by authors.

REFERENCES

1. Lim, H.J., Xue, H. and Wang, Y., (2020). Global trends in obesity. Handbook of Eating and Drinking: Interdisciplinary Perspectives, pp.1217-1235.
2. Omer, T., (2020). The causes of obesity: an in-depth review. Adv Obes Weight Manag Control, 10 (3), pp.90-94.
3. P.H. Wilding, J. (2001), Causes of obesity. Pract Diab Int, 18: 288-292. https://doi.org/10.1002/pdi.277
4. Collaborators, G.B.D. and Ärnlöv, J., (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet, 396(10258), pp.1223-1249.
5. Phelps, N. H., Singleton, R. K., Zhou, B., Heap, R. A., Mishra, A., Bennett, J. E., ... & Barbagallo, C. M. (2024). Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. The Lancet, 403(10431), 1027-1050.
6. Shao, G. (2022, December). Comparison ofprediction of obesity status based on differentmachine learning approaches with differentfactor quantities. In International Conference on Biomedical and Intelligent Systems (IC-BIS 2022) (Vol. 12458, pp. 881-888). SPIE.
7. Aslanpour, D. (2023). Machine Learning-Based Assessment of Obesity: An Investigation of Model Performance and Feature Selection. *University of California, Los Angeles.*
8. Fei, Y., (2020). California Rental Price Prediction Using Machine Learning Algorithms. *University of California, Los Angeles.*
9. Pérusse, L., Chagnon, Y. C., Weisnagel, S. J., Rankinen, T., Snyder, E., Sands, J., & Bouchard, C. (2001). The human obesity gene map: the 2000 update. *Obesity Research*, *9*(2), 135-169.
10. 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.
11. Jindal, K., Baliyan, N., & Rana, P. S. (2018). Obesity prediction using ensemble machine learning approaches. In Recent Findings in Intelligent Computing Techniques: *Proceedings of the 5th ICACNI 2017*, Volume 2 (pp. 355-362). *Springer Singapore.*