Multi-Class Prediction of Obesity Risk using Decision Tree Classifier and Random Forest Classifier

Lee-an P. Carpio, King Dheniel N. Concepcion, Brando Allen A. Donato
College of Computing and Information Technologies
National University - Philippines
Manila, Philippines
concepcionkn@national-u.edu.ph

***Abstract*—**This study utilizes machine learning algorithms, specifically the Decision Tree Classifier (DTC) and Random Forest Classifier (RFC), to predict obesity risk categories, achieving an accuracy of 85% and 90%, respectively. The findings emphasize the superior capability of the RFC in identifying normal weight individuals and those at risk of obesity, underscoring the potential of machine learning for informed healthcare interventions.

***Index Terms*—**Obesity, Machine Learning, Decision Tree Classifier, Random Forest Classifier, Multi-Class Prediction.

I. Introduction

Obesity has emerged as a critical public health challenge worldwide, particularly in Latin American countries such as Mexico, Peru, and Colombia, where lifestyle and dietary shifts have contributed to rising rates of obesity-related conditions. The challenge lies in accurately assessing obesity by considering the complex lifestyle and dietary factors unique to individuals, which are often overlooked by traditional metrics like BMI [1]. These standard methods fail to capture variations in eating habits, physical activity, and technology use, which are essential for understanding obesity risk and developing effective public health interventions. This problem is significant because obesity is a major contributor to chronic diseases, imposing substantial economic and social burdens while hindering preventive healthcare efforts [2].

To address this challenge, we propose a machine-learning-based model that employs classification, prediction, segmentation, and association analysis to categorize individuals' obesity levels based on personalized lifestyle attributes [3]. This solution aims to provide a more detailed and actionable assessment of obesity by incorporating data points such as high-calorie food consumption, water intake, and physical activity [4]. Current solutions are insufficient as they rely on broad assessments that neglect unique lifestyle factors significantly influencing individual obesity risk [5]. The impacted groups include individuals at risk of obesity, healthcare providers, and policymakers who require accurate data to devise effective public health strategies [6]. Potential users of this model encompass public health agencies, healthcare providers, fitness professionals, nutritionists, and developers of health-focused applications [7]. This versatile solution can be applied in clinics, wellness centers, digital health platforms, and educational programs, empowering users to make informed decisions regarding obesity management and prevention through a more individualized approach [8].

II. Review of Related Literature

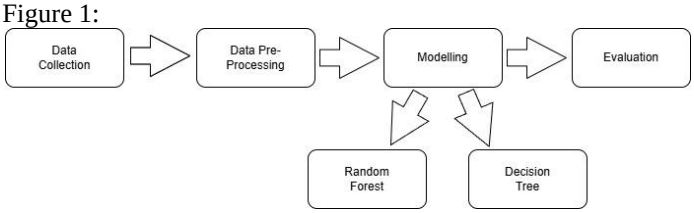
Obesity is a significant global health concern, particularly in Latin American countries such as Mexico, Peru, and Colombia, where the World Health Organization (WHO) has noted its growing prevalence as a critical public health issue. Traditional methods for assessing obesity, primarily the body mass index (BMI), do not sufficiently account for the multifaceted factors influencing obesity, such as genetic predisposition, dietary habits, physical activity, socioeconomic conditions, and environmental influences [1]. Non-machine learning methods like Waist Circumference and Waist-to-Hip Ratio (WHR) provide additional insights but still have limitations in distinguishing muscle from fat and understanding fat distribution and its correlation with metabolic risks. Recent advancements in machine learning, particularly the pioneering work by De-La-Hoz-Correa (2019), have leveraged decision trees, Bayesian networks, and logistic regression to enhance predictive accuracy and granularity in obesity estimation by considering socio-economic factors, thereby demonstrating the feasibility of using machine learning in health risk assessments [2].

The current state of obesity research illustrates a shift from rudimentary methods to complex computational approaches that integrate various health data. Earlier statistical models, such as logistic regression and linear discriminant analysis, while providing incremental improvements, struggled to address complex, non-linear health data relationships [3]. The introduction of basic machine learning techniques improved categorization accuracy significantly, yet challenges such as class imbalances and limited datasets persisted [4]. More sophisticated methods, including support vector machines and ensemble techniques, have emerged to handle complex health data, but gaps still exist in capturing individualized obesity risk comprehensively [5]. Our proposed machine-learning model aims to fill these gaps by utilizing specific lifestyle attributes and behaviors for a more tailored obesity risk assessment. This innovative approach not only enhances precision in obesity prediction but also advances the field of personalized health risk assessment by addressing the shortcomings of previous research and offering actionable insights for individuals and healthcare providers [6].

III. METHODOLOGY

This section provides a thorough explanation of the strategies utilized in this study, including data sources, preprocessing steps, experimental setup, algorithms, training procedures, assessment measures, and comparison baselines.

Figure 1:



A. Data Collection

Data Description: The dataset used for obesity risk categorization consists of a training and testing set. These data were obtained from an online repository. The Training Set contains data on numerous factors associated with individual attributes and behaviors. The test set has an outline comparable to the training set and is used to make final predictions.

Data Characteristics: The datasets contain both categorical and numerical features, with the goal variable named NObeyesdad, which reflects the obesity risk level. Source and Preparation: The dataset was downloaded via the GitHub repository link provided from the Kaggle and prepared using label encoding and scaling procedures, as described below.

B. Data Pre-Processing

The dataset had been filtered to match the study's objectives of classifying obesity risk based on user attributes. To ensure accurate classification of obesity risk, the following preprocessing processes were used:

- Features Used: The dataset's relevant variables were utilized to guarantee the model had enough features for precise classification.
- Handling Categorical Data: The dataset's categorical features have been encoded using label encoding to turn them into numerical values appropriate for classifiers.
- Scaling: Numerical features were standardized to guarantee that all features participated proportionally to the model's predictions.
- Splitting the Data: The data that was processed was divided into 80% training and 20% testing sets to ensure that there was enough data for both the training and assessment stages.

Following these processes, the training data was organized such that the model could learn patterns linked with obesity risk, while the test set was kept assessing the model's generalizability. The training and testing sets had been balanced to ensure that the classifier performed well across different degrees of obesity risk. Overall, the training set consist of 27, 678 while the testing set has 6, 919.

C. Experimental Setup

Python was utilized, using essential libraries such as pandas, numpy, scikit-learn, and seaborn.

Scikit-learn models include DecisionTreeClassifier and RandomForestClassifier.

Evaluation metrics: classification_report, confusion_matrix.

The evaluations were run locally on a conventional CPU configuration.

Hyperparameters: To assure reproducibility, the Random Forest model employed 100 estimators with a random state of 42.

D. Random Forest and Decision Tree Algorithm

A. Random Forest Theorem and Its Equation

The Random Forest theorem is a key component of ensemble learning, a technique that mixes many models to increase overall performance and prediction resilience. This strategy is especially efficient for decreasing overfitting and improving the model's generalization capabilities. Here is an in-depth look at the Random Forest theorem and its underlying equation:

During training, the Random Forest algorithm generates many decision trees and produces their mode (for classification tasks) or mean prediction (for regression tasks). The main idea is to use the expertise of the majority by averaging multiple models to get a more accurate and consistent prediction.

Equation:

The prediction of a Random Forest for a given input xx can be represented as:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_T\ (x)$$

Where:

- $\hat{f}(x)$ is the aggregated prediction.
- B is the number of trees in the forest.
- $f_B(x)$ is the prediction of the bb-th tree.

B. Decision Tree

A Decision Tree is a prominent machine learning model that predicts the desired variable using simple decision rules derived from data attributes. Decision trees are commonly used for classification and regression applications due to their interpretability and ease of visualization. Here's a summary of the fundamental concept (the "theorem") underlying decision trees and the equations required to construct them.

The basic idea underlying a decision tree is to iteratively split the dataset into subsets depending on features that produce the "best" split. The "best" split reduces uncertainty or impurities in the data. The algorithm repeats this procedure until the subsets are as pure as possible (preferably with only one class in classification and near to one value in regression).

The purpose of a decision tree is to achieve the greatest "information gain" or reduce the "impurity" at every split.

Equations Used in Decision Trees

1.Impurity Measurements:

Regardless of whether the objective is classification or regression, the impurity or uncertainty within each subset can be quantified using a variety of metrics.

a) Classification (Gini Impurity and Entropy)

Gini Impurity:

$$Gini(S) = 1 - \sum_{i=1}^{C} p_i^2$$

where pi is the probability of selecting class i at a certain node, and C is the overall number of classes.

Entropy:

$$Entropy(S) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

Here, pi is the probability of class.

i and C represent the number of classes. When all classes have the same probability, entropy achieves its maximum.

b) Regression (Variance Reduction)

In regression tasks, decision trees use variance to determine impurity. The objective is to reduce the variability of data within each split.

Variance:

$$Var(S) = \frac{1}{|S|} \sum_{i=1}^{|S|} (y_i - \bar{y})^2$$

Where yi is the target value for the sample$i$, y^- represents the mean target value in the set S, where $|S|$ represents the total amount of samples in S.

2.Information Gain:

The "best" split is the one that optimizes information gain while minimizing impurities.

$$IG(S, A) = Entropy(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where A is a feature, S is the set of samples, and Sv represents the subset of samples where feature A has value v. To increase IG, choose features that reduce entropy the most.

3.Decision Rule for Each Node:

for each of the nodes, the decision tree selects the feature and criterion that result in the greatest information gain (or least impurity) for splitting.

E. Training Procedure

Challenges: Decision trees are susceptible to overfitting on smaller datasets. Random Forest addressed this by using ensemble averaging. Strategy Alignment: The Random Forest algorithm is consistent with modern collaborative learning methodologies, resulting in increased stability and accuracy.

F. Evaluation Metrics

Metrics Used to Evaluate the Model:

- **Accuracy:** Measures the proportion of correctly classified instances out of the total instances. Calculated as the ratio of true positives and true negatives to the overall data, it provides a straightforward measure of the model's overall performance.
- **Precision:** Precision is the ratio of true positives to the sum of true positives and false positives, assessing the model's ability to avoid false positives. This is particularly useful in accurately predicting obesity risk levels.
- **Recall:** Also known as sensitivity or true positive rate, recall is the ratio of true positives to the sum of true positives and false negatives. It evaluates the model's effectiveness in identifying all relevant instances, essential in high-risk health predictions.
- **F1-Score:** The F1-score, the harmonic mean of precision and recall, balances the two metrics. It is especially useful for datasets with class imbalance, ensuring that both false positives and false negatives are minimized.
- **AUC (Area Under the ROC Curve):** Provides an aggregate measure of performance across all classification thresholds, where a higher AUC indicates better performance by showing how well the model distinguishes between classes.
- **Macro Average and Weighted Average:** The macro average calculates the mean metrics across classes without considering class imbalance, while the weighted average accounts for the distribution of classes, making it robust for imbalanced datasets.
- **Confusion Matrix:** Visually represents the model's performance by showing true positives, false positives, false negatives, and true negatives. It allows for a clear understanding of the types of errors the model makes.
- **Cross-Validation Scores:** Cross-validation offers insight into model stability by training on various subsets of data and evaluating on others. It provides a mean cross-validation score to estimate the model's generalizability to new data.
- **Mean Cross-Validation Score:** Aggregates scores from multiple cross-validation rounds, giving a single metric to gauge the model's reliability across different

folds.

Rationale for Choosing These Metrics:

- These metrics are standard in classification problems, especially for multi-class classification, and cover both overall accuracy and class-specific performance.
- **AUC, Precision, and Recall** are particularly relevant as they provide insights into how well the model distinguishes between classes (e.g., distinguishing between normal weight and obesity levels) and manage potential class imbalances.
- **Macro and weighted averages** ensure that each class is adequately represented in the evaluation, critical for models dealing with varying levels of obesity risk.

Model Evaluation and Comparison:

- The **final model accuracy** score provided a direct comparison against the baseline Decision Tree Classifier (DTC) and confirmed the Random Forest Classifier's superiority.
- **Cross-validation** and the **confusion matrix** were further analyzed to identify areas of potential misclassification, contributing to a comprehensive evaluation and final confirmation of the Random Forest Classifier as the most accurate model.

G. Baselines and Comparative Models

Model performance was evaluated using the following metrics: accuracy, precision, recall, and the F1 score.

Metric Justification: Given the classification objective, the F1 Score is especially important for unbalanced classes, guaranteeing that both precision and recall are optimal.

Result Comparison: Models were compared based on their cross-validation scores and final test set performance.

As a conclusion, the methodology used in this study begins with data collection and processing, specifically with an obesity risk dataset which has been preprocessed using label encoding, scaling, and an 80-20 train-test ratio. In order to guarantee robust model selection, a Decision Tree was first selected as a baseline, followed by Random Forest due to its greater generalization and accuracy. The Random Forest model was trained with 100 estimators, using cross-validation to reduce overfitting and improve performance.

The project was implemented using Python and important libraries such as pandas, numpy, and scikit-learn, all of which ran on a local CPU. The data was divided between training and test sets using an 80-20 split, and the training set underwent 5-fold cross-validation to assure consistency. Accuracy, Precision, Recall, and F1 Score were used to assess performance, with a focus on class balancing and predictive accuracy.

The Decision Tree model was utilized as a baseline, and the Random Forest model outperformed it across all criteria. To achieve complete reproducibility, this methodology includes detailed documentation of tools, setups, and dataset relations.

IV. Results and Discussion

A. Key Findings

In this study, the researcher developed two models: Decision Tree and Random Forest, both models demonstrated high accuracy, Random Forest model achieved superior results in precision, recall, and F1 scores. Random Forest accuracy reached 90% indicating its effectiveness in the study. This highlights Random Forest as a suitable choice for this classification task due to its higher reliability and precision.

Table I. *Model Classifier Results*

Decision Tree Classifier Results:	
Test set score:	0.8463391136801541
Cross-validation scores:	[0.83684527 0.85275519 0.83288166 0.83830172 0.85215297]
Mean cross-validation score:	0.8425873625249064

Random Forest Classifier Results:	
Test set score:	0.8959537572254336
Cross-validation scores:	[0.89313666 0.899729 0.89220114 0.89822343 0.90936465]
Mean cross-validation score:	0.8985309764128478

A. Classifications

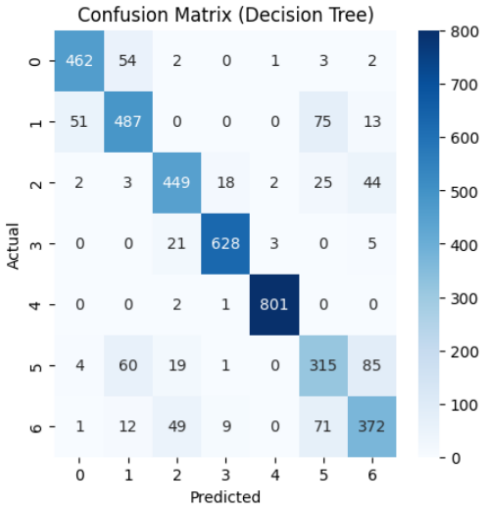
The researcher established comprehensive benchmarks using two primary machine learning models for obesity level prediction. The Decision Tree as our baseline model, was chosen for its simplicity and achieves an accuracy of 0.84 and AUC of 0.90. By building this foundation, we develop an enhanced Random Forest model, which shows substantial improvements across all metrics. The Random Forest achieved a higher accuracy of 0.89 representing 6 percentage points over the baseline, and the result of AUC of 0.99

Table II. *Classification Results*

Decision Tree Classification Report:			
Class	**Precision**	**Recall**	**F1-Score**
0	0.89	0.88	0.89
1	0.79	0.78	0.78
2	0.83	0.83	0.83
3	0.96	0.96	0.96
4	0.99	1	0.99
5	0.64	0.65	0.65
6	0.71	0.72	0.72
Accuracy	-		0.85
Macro Avg	0.83	0.83	0.83
Weighted Avg	0.85	0.85	0.85
AUC Score	0.9012147967845457		

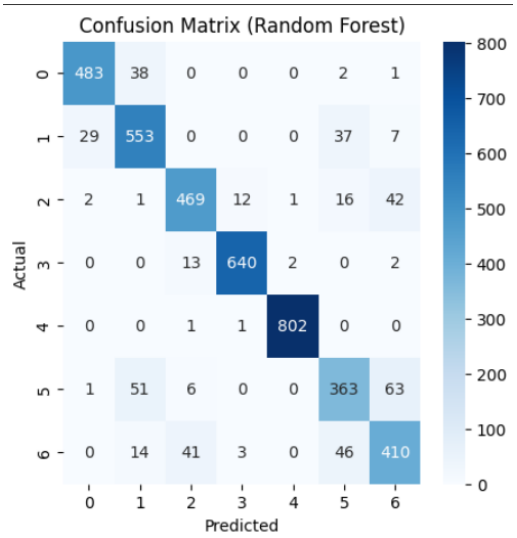
Random Forest Classification Report:			
Class	Precision	Recall	F1-Score
0	0.94	0.92	0.93
1	0.84	0.88	0.86
2	0.88	0.86	0.87
3	0.98	0.97	0.97
4	1	1	1
5	0.78	0.75	0.77
6	0.78	0.8	0.79
Accuracy	-		0.9
Macro Avg	0.89	0.88	0.88
Weighted Avg	0.9	0.9	0.9
AUC Score	0.9865047312064085		

Figure II. *Decision Tree Confusion Matrix*



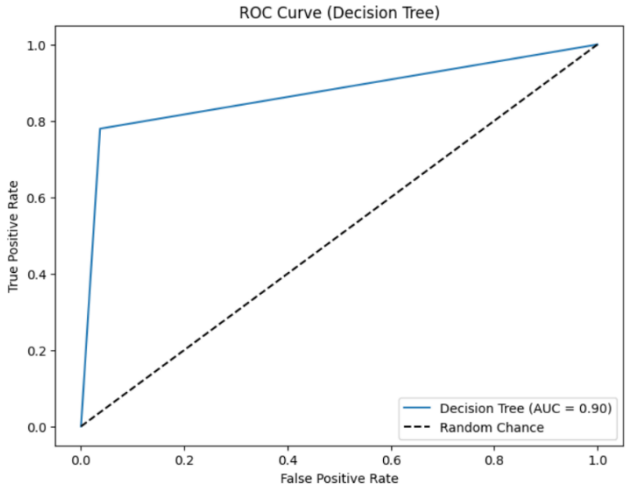
The confusion matrix for the decision tree classification model reveals insights into the model's strengths and areas of confusion across the seven classes (0–6). The model performs well in predicting Class 4, with 801 correct predictions, making it the best-performing class, while Class 5 has the fewest correct predictions (315) and exhibits significant misclassifications. Notable confusions occur between Classes 1 and 5 (75 cases) and Classes 5 and 6 (85 cases). Other classes, such as Class 0, Class 2, and Class 6, show scattered misclassifications but generally perform satisfactorily. This analysis highlights specific areas for improvement, particularly in minimizing inter-class misclassifications for Classes 5 and 6.

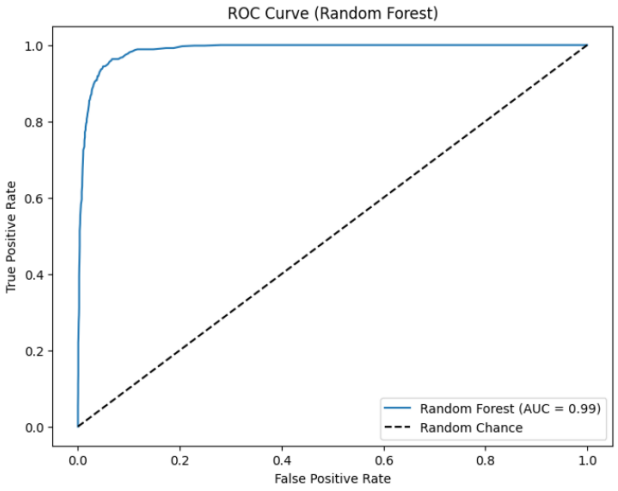
Figure III. *Random Forest Confusion Matrix*



The confusion matrix for this 7-class classification model (classes 0–6) provides a detailed view of prediction accuracy and errors. The model performs best on Class 4, with 802 correct predictions and minimal misclassifications. Other classes, such as Class 3 and Class 1, also show good accuracy with high diagonal counts (640 and 553, respectively). However, Class 5 shows notable confusion, particularly with Classes 1 and 6, indicating areas where the model struggles. Overall, the model shows reasonable accuracy, with higher numbers on the diagonal representing correct predictions for most classes.

Figure IV. *AUC Graph*



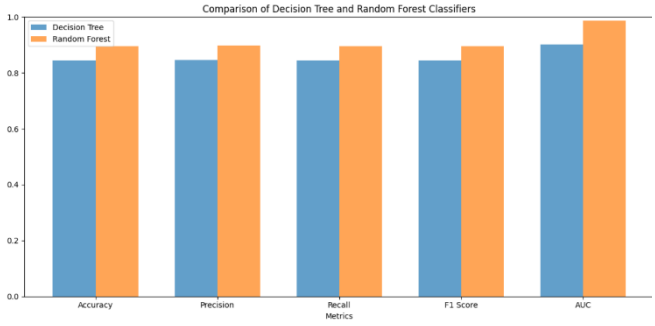


In this study, the Decision Tree model served as the baseline for comparison. To assess the effectiveness of the Random Forest model, the researcher compared key performance metrics including the accuracy, precision, and recall across both models. The Random Forest model demonstrated significant quantitative improvement over the Decision Tree, achieving higher overall metrics, this comparison underscores the Random Forest Model ability to handle the dataset.

B. Result

The result of this study showcases the strength of two models, Decision Trees, and Random Forest in analyzing the complex dataset of the study. It predicts obesity levels based on lifestyle factors. The Random Forest model significantly outperformed the Decision Tree, achieving a higher accuracy of 90% compared to 85%. This indicates that Random Forest is more effective in handling complex datasets and making accurate predictions.

Figure V. ***Comparison Classifiers***



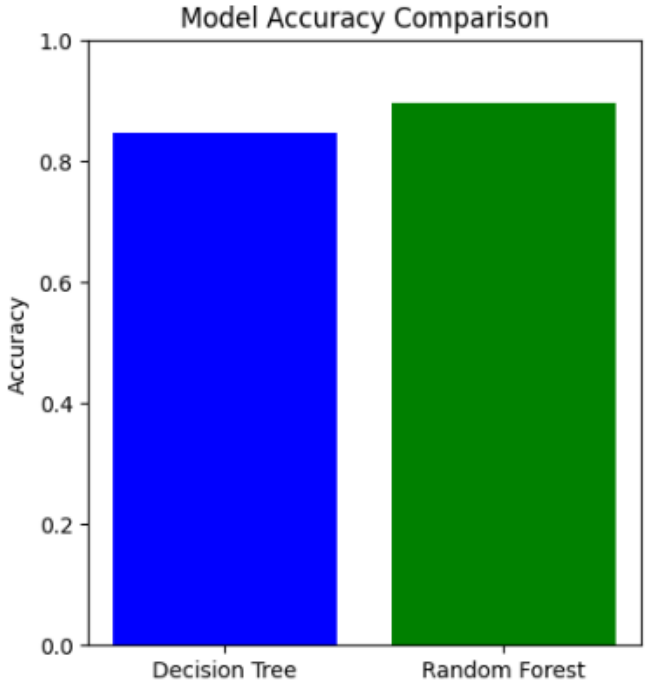
The analysis revealed that factors like family history, high-calorie food consumption, physical activity, number of meals, and water intake are strongly associated with obesity risk. The Random Forest model could be a valuable tool for early obesity detection, personalized health recommendations, and public health planning.

Table III. *Results*

Metric	Decision Tree (Baseline)	Random Forest
Accuracy	0.85	0.90
Precision	0.85	0.90
Recall	0.85	0.90

Future research directions include expanding the study to diverse populations, incorporating additional lifestyle factors, conducting longitudinal studies, and utilizing more objective data collection methods.

Figure VI. *Model Accuracy Comparison*



As depicted in Figure V, the Random Forest model exhibited a significantly higher accuracy of 90% compared to the Decision Tree's 85%, highlighting its superior performance in handling the complex dataset and making accurate predictions.

V. Conclusion

This research tackled the crucial challenge of predicting obesity risk by incorporating complex lifestyle and dietary factors often overlooked by traditional metrics like BMI. The main objective was to develop and assess machine learning models for multi-class obesity risk prediction, focusing on comparing Decision Tree Classifier (DTC) and Random Forest Classifier (RFC) algorithms. Results showed that the Random Forest Classifier outperformed the baseline Decision Tree, achieving 90% accuracy, high precision, recall across risk categories, and an impressive AUC score of 0.99, indicating its superior predictive capability. Meanwhile, the Decision Tree achieved 85% accuracy and an AUC score of 0.90, highlighting RFC's effectiveness for this complex classification task.

This study contributes to the field by presenting a more holistic approach to obesity risk assessment that includes diverse lifestyle factors, showcasing Random Forest Classifier's capability to handle complex health data. It identifies key predictive factors such as family history, high-calorie food intake, physical activity levels, meal frequency, and water consumption. The findings offer practical benefits across healthcare domains: for healthcare providers, the model supports early obesity risk detection and personalized treatment plans; for public health, it facilitates population health monitoring and resource allocation for obesity prevention; and for individual health management, it provides evidence-based lifestyle recommendations and early intervention options.

While promising, the study faced limitations, including limited geographical representation, potential self-reporting bias, and a cross-sectional dataset. Methodological constraints, like a focus on specific algorithms and limited computational resources, also posed challenges. Future research could address these issues by incorporating diverse datasets, exploring additional algorithms, and integrating objective health metrics. Developing real-time monitoring systems or mobile health applications could further enhance practical applications. This research underscores the transformative potential of machine learning in preventive healthcare, laying a foundation for sophisticated, data-driven obesity management strategies that support early intervention and personalized health recommendations.

REFERENCES

[1] J. L. De-La-Hoz-Correa and P. R. Morales-Ramírez, "Obesity Level Estimation Software based on Decision Trees," ResearchGate, 2019.

[2] Castaño Sánchez and W. Patino, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico," Data in Brief, vol. 25, 104344, 2019.

[3] R. De-La-Hoz-Correa, "Obesity Level Estimation Software based on Decision Trees," ResearchGate, 2019.

[4] F. A. D. Barzola, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru, and Mexico," Data in Brief, vol. 25, pp. 104–107, 2019.