Research Paper
https://doi.org/10.55525/tjst.1572382

Machine Learning Models for Accurate Prediction of Obesity: A Data-Driven Approach

Ali DEĞİRMENCİ[1*]
[1] Department of Electrical and Electronics Engineering, Faculty of Engineering and Natural Sciences, Ankara Yıldırım Beyazıt University, Ankara, Türkiye
[*1] alidegirmenci@aybu.edu.tr

(Geliş/Received: 23/10/2024; Kabul/Accepted: 03/12/2024)

Abstract: The number of people affected by obesity is rising steadily. Diagnosing obesity is crucial due to its harmful impacts on human health and it has become one of the world's most important global health concerns. Therefore, it is crucial to develop methods that can enable early prediction of obesity risk and aid in mitigating the increasing prevalence of obesity. In the literature, some methods rely solely on Body Mass Index (BMI) for the prediction and classification of obesity may result in inaccurate outcomes. Additionally, more accurate predictions can be performed by developing machine learning models that incorporate additional factors such as individuals' lifestyle and dietary habits, alongside height and weight used in BMI calculations. In this study, the potential of three different machine learning methods (naive Bayes, decision tree, and Random Forest (RF)) in predicting obesity levels were investigated. The best performance among the compared methods was obtained with RF (accuracy=0.8892, macro average F1-score=0.8618, Macro Average Precision (MAP)=0.8350, Macro Average Recall (MAR)=0.9122,). In addition, feature selection was also performed to determine the features that are significant for the estimation of the obesity level. According to the experimental results with feature selection, the RF method resulted in the highest score (accuracy=0.9236, MAP=0.9232, MAR=0.9358, macro average F1-score=0.9269) with fewer features. The results demonstrate that the performance of machine learning models on the same dataset can be enhanced through detailed hyperparameter tuning. Furthermore, applying feature selection can improve performance by mitigating the adverse effects of irrelevant or redundant features that may degrade the model's effectiveness.

Key words: Obesity, machine learning, feature selection, mutual information.

Obezitenin Doğru Tahmini için Makine Öğrenimi Modelleri: Veri Odaklı Yaklaşım

Öz: Obezitenin insan sağlığı üzerindeki zararlı etkileri ve obeziteden etkilenen bireylerin sayısı giderek artışı nedeniyle bu sorunun teşhis edilmesi büyük bir önem taşımaktadır. Obezitenin yaygınlaşması küresel sağlık açısından en önemli sorunlardan biri haline gelmesine yol açmıştır. Bu nedenle, obezite riskinin erken tespitini sağlayacak, ayrıca obezitenin artan yaygınlığını azaltmaya yardımcı olacak yöntemlerin geliştirilmesi elzemdir. Obezitenin öngörülmesi ve sınıflandırılması için yalnızca Beden Kitle İndeksine (BKİ) güvenmek hatalı sonuçlara yol açabilir. BKİ hesaplamalarında kullanılan boy ve kilonun yanı sıra bireylerin yaşam tarzı ve beslenme alışkanlıkları gibi ek faktörleri de içeren makine öğrenimi modelleri geliştirilerek daha doğru tahminler elde edilebilir. Bu çalışmada, üç farklı makine öğrenimi yönteminin (naive Bayes, karar ağacı ve Rasgele Orman (RF)) obezite seviyelerini tahmin etme potansiyeli araştırılmıştır. Karşılaştırılan yöntemler arasında en iyi performans RF ile elde edilmiştir (doğruluk=0,8892, makro ortalama F1-skor=0,8618, Makro Ortalama Kesinlik (MAP)=0,8350, Makro Ortalama Duyarlılık (MAR)=0,9122). Ayrıca, obezite seviyesini tahmin etmede etkili olan öznitelikleri belirlemek için öznitelik seçimi de yapılmıştır. Öznitelik seçimi ile elde edilen deneysel sonuçlara göre, RF yöntemi daha az öznitelik ile en yüksek skoru (doğruluk=0,9236, MAP=0,9232, MAR=0,9358, makro ortalama F1-skor=0,9269) elde etmiştir. Sonuçlar, makine öğrenimi modellerinin aynı veri kümesi üzerindeki performansının ayrıntılı hiperparametre ayarlamasıyla artırılabileceğini göstermektedir. Ayrıca, öznitelik seçimi uygulamak, modelin etkinliğini azaltabilecek ilgisiz veya gereksiz özniteliklerin olumsuz etkilerini azaltarak performansı artırabilir.

Anahtar kelimeler: Obezite, makine öğrenmesi, öznitelik seçimi, karşılıklı bilgi.

1. Introduction

The rising prevalence of obesity has emerged as a major concern in global public health. Obesity is a complex, multifactorial health issue that can affect individuals of any age, regardless of location, ethnicity, or socioeconomic status, and has thus become a global epidemic [1]. In recent years, factors such as easy transportation, decreased physical activity, long screen time, increased consumption of processed foods, as well as sedentary lifestyles have led to an increase in obesity. Moreover, it is well known phenomenon that obesity is linked to various health conditions, including heart disease, type 2 diabetes, and certain cancers (such as colorectal, endometrial, liver, pancreatic, and kidney cancers). It also increases the risk of surgical procedures, metabolic abnormalities, joint problems, and other chronic diseases [2]. There is a strong connection between obesity and diabetes (especially type 2 diabetes). Obesity causes insulin resistance, making it harder for the body to use insulin effectively. When

[*] Corresponding author: alidegirmenci@aybu.edu.tr. ORCID Number of authors: [1] 0000-0001-9727-8559

this happens, blood sugar levels rise, and the pancreas tries to produce more insulin to keep blood sugar levels under control. Over time, the pancreas cannot sustain this excessive production and Type 2 diabetes can develop. Recently, studies focusing on determining blood glucose through non-invasive methods have gained significant attention [3]. Obesity accounts for a significant proportion of health expenditure and imposes a substantial burden on society. Therefore, both direct and indirect costs of obesity have an important place in the health system expenditures of countries. For this reason, measures should be taken to prevent obesity. This can be possible by raising public awareness of the factors that cause obesity and the health problems caused by obesity and taking action accordingly [4,5].

The World Health Organization (WHO) defines obesity as excessive fat accumulation in the body to the extent that it impairs health. Obesity arises from the fact that the energy intake from daily food is more than the energy expended, and this excess is stored as fat in the body. Body Mass Index (BMI) is a straightforward and widely adopted tool employed to assess and classify obesity based on a person's height and weight [6,7]. BMI is defined as in Equation (1).

$$BMI = \frac{Weight\ in\ kilograms}{Height\ in\ meters^2} \tag{1}$$

BMI categories defined by the WHO are given in Table 1. According to Table 1, weight status is categorized into four main groups, with obesity further classified into three subcategories based on BMI. Within these categories, the risk of developing health problems increases, except for individuals with normal weight, and the severity of these risks escalates with the level of obesity.

Table 1. BMI categories based on WHO.

Category	**Obesity class**	$BMI\ (kg/m^2)$
Underweight	-	< 18.5
Normal weight	-	$18.5 - 24.9$
Overweight	-	$25.0 - 29.9$
Obese	Class I	$30.0 - 34.9$
	Class II	$35.0 - 39.9$
	Class III	> 40.0

Although BMI is widely adopted as a measure of obesity, it is insufficient as a stand-alone metric due to its inherent limitations and inaccuracies. BMI fails to discriminate between body fat and muscle mass. Consequently, individuals such as weightlifters and athletes can be categorized as overweight or obese, while those with low muscle mass but high fat levels may be misclassified as healthy. BMI also does not take into account other health indicators such as cholesterol levels, blood pressure, and metabolic health. Body composition can vary between different ethnic groups, age ranges, and genders, so using the same index for everyone can lead to inaccurate assessments. The BMI also fails to consider body fat distribution; studies have shown that fat around the abdomen (visceral fat) has been shown to better predict health risks than fat stored elsewhere [8].

Machine learning-supported approaches have been widely adopted in many fields, such as outlier detection [9,10], medicine [11-13], and biology [14]. Machine learning based approaches have also been adopted in obesity prediction. Previous reports showed that different hypotheses and models were developed for the estimation and classification of obesity using various machine learning techniques. For instance, Cheng et al. employed a recurrent neural network-based model, specifically Long Short-Term Memory (LSTM), to estimate BMI in children [15]. The data set was obtained from the Obesity Prediction in Early Life (OPEL) database. It consists of children aged 0 to 4 years, 2, 3, 5, and 8 clinic visits according to the electronic health record. According to the findings, five visits were adequate for accurate forecasts, the performance results of the LSTM model showed a mean absolute error of 0.98 and R^2 of 0.72. Solomon et al. presented a majority voting-based ensemble learning model that includes an eXtreme Gradient Boosting (XGBoost), a gradient-boosting classifier, and a Multi-Layer Perceptron (MLP) to predict and classify obesity [16]. The dataset has 17 features, including eating habits and physical conditions, from 2111 people in Colombia, Mexico, and Peru. To demonstrate the effectiveness of the model, it was compared with different machine learning algorithms (Naive Bayes (NB), XGBoost, Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), K-Nearest Neighbor (KNN), MLP, and Random Forest (RF)). While their model achieved the highest accuracy of 97.16%, the closest result was obtained using XGBoost at 96.37%. Another study using the same dataset was conducted by Kaur et al. [17]. Six different machine learning methods (GB, Bagging Meta-Estimator (BME), XGBoost, RF, SVM, and KNN) were used. Different train/test ratios (90:10, 80:20, 70:30, 60:40) were examined. The highest accuracy in each train/test ratio was achieved by GB at 90:10 ratio (98.11%), GB and XGboost at 80:20 ratio (97.87%), and XGboost at 70:30 ratio (97.79%). In addition, meal recommendations were made according to calorie and macronutrient needs with the nearest neighbor learning method. Wang et al. used 9 different machine learning methods (Logistic Regression (LR), NB, KNN, DT, SVM, Light Gradient Boosting Machine (LGBM), RF, GB Machine (GBM), and XGBoost) for overweight or obesity risk prediction in Chinese preschool-aged children [18]. The dataset includes a total of 9478 children, 1250 of whom were overweight or obese. With the training-test ratio of 6:4, the SVM algorithm achieved the highest accuracy (0.9457). The top 5 most influential features were identified through χ2-based Scikit-learn feature selection and Shapley additive explanation. Liu et al. utilized different machine learning algorithms to measure the relationship between obesity status and BMI values [19]. Gut microbiota metagenomics and phenotype information data were gathered from 2262 Chinese volunteers to investigate microbiota-obesity interaction. The best prediction accuracy of BMI groups among Gradient Boosting DT (GBDT), RF, SVM, logistic regression methods was obtained with SVM (0.716). Wong et al. compared the performance of 4 methods, XGBoost, RF, LR, and SVM, to identify overweight or obesity status among working adults in Malaysia [20]. The dataset consists of 16,860 individuals, of which 7048 were overweight or obese. 70% of the data is utilized for training and the remaining 30% for testing, with the data being randomly split. The Area Under the Receiver Operating Characteristic (ROC-AUC) curve results for the methods are XGBoost = 0.81, RF and SVM = 0.80, and LR = 0.78. The results indicate that the performances of the compared models are not significantly different from each other. Calderón-Díaz et al. classified Chilean youth into two categories: normal weight and overweight/obese [21]. The dataset consists of 13 biomedical and 8 lipid features obtained from 40 university students between the ages of 20-30. XGBoost was applied to this dataset and a ROC-AUC score of 0.818 was obtained. Köklü and Sulak categorized the Turkish people as underweight, normal, overweight, and obese [22]. The dataset in their study consists of 14 features obtained from 1610 individuals. They compared the performance of Artificial Neural Network (ANN), KNN, RF, and SVM methods. The accuracies of the compared methods are 74.96% for ANN, 74.03% for SVM, 80.62% for KNN, and 87.82% for RF.

Predicting the individuals' obesity is of great interest because it affects the quality of life. As can be seen in the literature review, studies on obesity level prediction using machine learning methods are generally carried out on a country level and/or for specific age groups. Studies on predicting obesity in Turkey using machine learning are quite limited. It is of great importance to determine the factors affecting obesity on a national level. In this study, obesity levels were predicted by using three different machine learning methods: random forest, naive Bayes, and decision tree. To achieve the best performance among these methods, hyperparameter optimization was performed using the brute force grid search method, covering a specified range of method-specific hyperparameters. Then, most prominent features that have the effect of the estimation are analyzed with the Mutual Information (MI) feature selection method. To train the machine learning methods compared from the hyperparameters that achieve the best performance using all samples in the data set, different subsets of the data were created from the highest to the lowest according to the MI score. In this process, one feature was added at a time and the performance of the algorithms was assessed with each subset. In this way, the subset with the highest success rate was determined by using the minimum number of attributes. According to the experimental results, the best results obtained with the RF method using the entire data set are accuracy: 0.8892, macro average F1-score= 0.8618, Macro Average Precision (MAP)=0.8350, and Macro Average Recall (MAR)=0.9122. Using feature selection, the highest scores were obtained with 9 features and the performance scores were accuracy=0.9236, macro average F1-score=0.9269, MAP=0.9232, and MAR=0.9358. These results highlight that feature selection can improve the performance of machine learning algorithms.

The structure of the study is defined as follows. Section 2 introduces the dataset used in the study, provides a description of the machine learning methods compared and the feature selection method. Section 3 presents the performance of the machine learning algorithms, the MI scores of the features in the dataset, and the analysis of results obtained using fewer features based on their MI scores. Section 4 provides concluding remarks and outlines future directions for the study.

2. Materials and Methods

This study conducts a comparative analysis of three different machine learning methods used to predict obesity levels from an online questionnaire. The machine learning methods employed are NB, DT, and RF. In these methods, hyperparameter tuning was performed with the brute force grid search method to achieve the best performance. Then, the importance of each feature in the dataset was determined through MI. Using the best-performing method and its optimized hyperparameters, features were sequentially added from the highest to the lowest MI score. The positive and negative effects of adding each feature on the model's performance were then examined. In all these analyses, k-fold cross-validation was employed to increase the reliability of the machine learning algorithms. The structure of the study is presented in Figure 1. The subsequent subsections outline the employed machine learning methods and feature selection technique, offering a thorough explanation of the methodology used in the study.

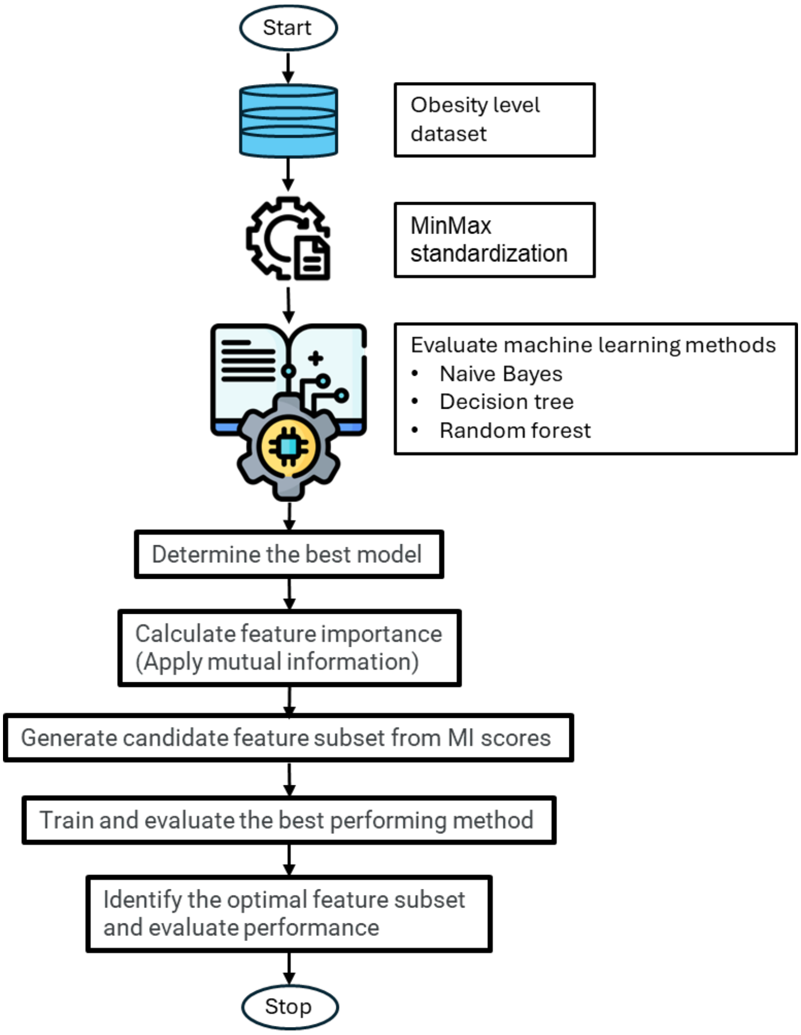

Start
Obesity level dataset
MinMax standardization
Evaluate machine learning methods
• Naive Bayes
• Decision tree
• Random forest
Determine the best model
Calculate feature importance (Apply mutual information)
Generate candidate feature subset from MI scores
Train and evaluate the best performing method
Identify the optimal feature subset and evaluate performance
Stop

Figure 1. The flowchart of the study.

2.1. Data set

The data set employed in this study was acquired from a Kaggle repository named "Obesity" [22,23]. It was gathered through an online questionnaire from individuals residing in Turkey. The purpose of the data set was to determine the obesity levels of individuals based on the characteristics identified as influencing obesity in literature. Individuals were divided into four different classes: underweight, normal, overweight, and obese. The dataset consists of 1610 instances and 14 features. The dataset exhibits a skewed class distribution, as the proportions of each class are unequal. The class distributions are as follows: underweight (4.5%), normal (40.8%), overweight (36.8%), and obesity (17.8%). Details of the features, including brief descriptions, are given in Table 2.

Table 2. Explanations of the features in the Obesity data set.

Feature	Explanation	Measurement	Range
Age	Age of individual	Years	[18,...,54]
Sex	Male / Female	Boolean	0,1
Height	Height of individual	cm	[150,...193]
Overweight/Obese family	Overweight/Obese history in the family	Boolean	0,1
Fastfood	Fast food consumption	Boolean	0,1
Vegetable consumption	Frequency of vegetable consumption	Categorical	Rarely, Sometimes, Always
Main meals	Number of daily main meals	Categorical	1-2, 3, 3+
Interval meal	Intermeal food consumption	Categorical	Rarely, Sometimes, Usually, Always
Smoking	Smoking habit	Boolean	0,1
Liquid consumption	Daily liquid consumption	Categorical	<1 liter 1-2 liters > 2 liters
Calorie calculation	Calculation of daily calorie consumption	Boolean	0,1
Physical activity	Physical activity per week	Categorical	No activity 1-2 days 3-4 days 5-6 days 6+ days
Screen exposure	Duration of screen exposure	Categorical	0 - 2 hours 3 - 5 hours > 5 hours
Transportation type	Mode of transportation utilized	Categorical	Automobile, Motorbike, Bike, Public transportation, Walking

2.2. Naive Bayes (NB)

The NB classifier is a probabilistic classifier that employs Bayes' theorem to classify samples. In this method, the assumption is made that the features in the dataset are independent of each other. Bayes' rule is based on the fundamental concept of conditional probability, enabling users to calculate the probability of event *C* occurring given that event *X* has also occurred *p(C|X)* [24]. It is defined as in Equation (2).

$$p(C|X) = \frac{p(X|C)p(C)}{p(X)} \tag{2}$$

where *p(X|C)* is probability of *X* occurring, given *C* has occurred, *p(C)* equals probability of *C* occurring, and *p(X)* is probability of *X* occurring.

Assume $X = \{x_1, x_2, \dots, x_n\}$ to be a set of feature vectors of a new sample to be classified, where $x_1, x_2, \dots, x_n$ corresponds to the features of the sample, and $C = \{c_1, c_2, \dots, c_M\}$ is the target of possible classes. Here, *M* denotes the number of classes in the data set. Equation (2) can be expressed for *n* number of events x_i to be occurred, as shown in Equation (3).

$$p(c_j|x_1, x_2, \dots, x_n) = \frac{p(x_1, x_2, \dots, x_n|c_k)p(c_j)}{p(x_1, x_2, \dots, x_n)} \tag{3}$$

where c_j is the j^{th} class. Posterior probability $p(x_i|y)$ estimation increases model complexity and requires to need of excessive training data. Hence, in NB, it is assumed that all the features are conditionally independent. Additionally, $p(x_1, x_2, \dots, x_n)$ is constant for a sample and equal across classes in the data set. NB classifier can be defined as in Equation (4).

$$p(c_j|x_1, x_2, \dots, x_n) \propto \prod_{i=1}^{n} p(x_i|c_j)\, p(c_j) \tag{4}$$

The classification of the sample can be defined as in Equation (5).

$$\hat{y}_i = \underset{c_j \in C}{\arg\max}\left[p(c_j)\prod_{i=1}^{n} p(x_i|c_j)\right] \tag{5}$$

where y_i is the estimated class with maximum a posteriori probability, and $p(x_i|c_j)$ is the distribution of the i^{th} attribute for the given class c_j.

2.3. Decision tree (DT)

A DT is a tree-like structured model and a nonparametric method, meaning it does not rely on any assumptions about the distribution of the input data. It is a top-down approach that predicts the class of the unseen sample using decision rules. DT consists of root node, internal nodes, and leaf nodes. The root node is the top node of the tree where branching begins. Internal nodes reside between the root node and the leaf nodes, representing the conditions that determine how the tree splits into branches. Leaf nodes are also known as terminal nodes and determine the class of the query instance. In splitting, the aim is to decrease the impurity (or uncertainty) in the dataset corresponding to the class at a later stage. Commonly used techniques to select the splitting criteria include information gain and the Gini index [25].

To determine the optimal cut-off feature, features that have a lower Gini index are preferred in the DT method. The Gini index is computed as shown in Equation (6).

$$Gini\ \ Index = 1 - \sum_{i=1}^{n} p_i^2 \tag{6}$$

where p_i is the probability of the i^{th} class in the dataset and n is the number of classes in the data set. Sample schematic of the DT is shown in Figure 2.

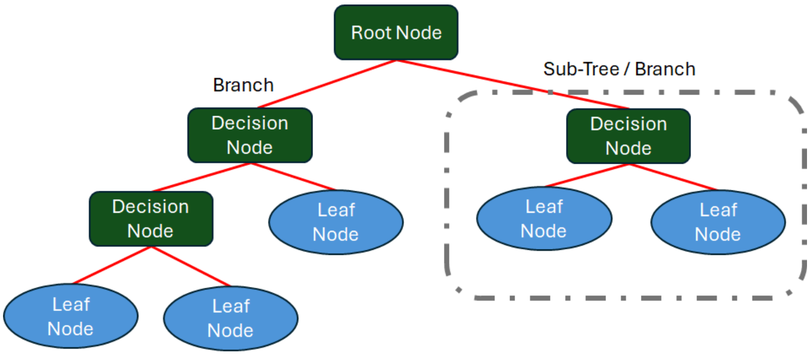


Figure 2. Schematic of the decision tree.

2.4. Random forest (RF)

RF is an ensemble learning algorithm that combines the concept of bootstrapping and aggregation, abbreviated as bagging. The intuition behind ensemble learning algorithms is that by combining a group of models, the performance and robustness of predictions can be improved compared to individual models. This is because a group of models, also defined as weak learners, can produce a powerful model that outperforms any individual model when aggregated. In bootstrapping, a new data set is created with replacement and samples are randomly selected from the original data set. The size of the training data set and the bootstrapped data set are equal. In aggregation, model predictions are averaged in regression problems, whereas in classification problems, the final decision is determined by majority voting. It is defined as in Equation (7).

$$\hat{y} = \underset{m \in \{1,2,\ldots,M\}}{\arg\max} \sum_{i=1}^{P} I(C_i = m) \tag{7}$$

where M is the number of classes, C_i equals the prediction of the i^{th} classifier, P is the number of classifiers, $\hat{y}$ denotes the final class prediction.

The DT method is used as a weak learner in RF. Each tree model in the RF method is created with bootstrapped data sets. In RF, randomness is added to the bagging. To grow a tree in RF, instead of using the best split of features in the DT method, it employs the best split among the random subset of features in the division of each node. Although this process weakens the strength of each tree, it also reduces the correlation and generalization error between the trees [26]. In the RF classification, the final prediction on unseen samples is determined by a majority vote of the prediction made by each of the tree models built from the bootstrapped data. The workflow of RF can be visualized in Figure 3. Nodes are split using a random subset of features, and branches terminate at leaf nodes, which contribute to the final decision based on the path followed in each tree.

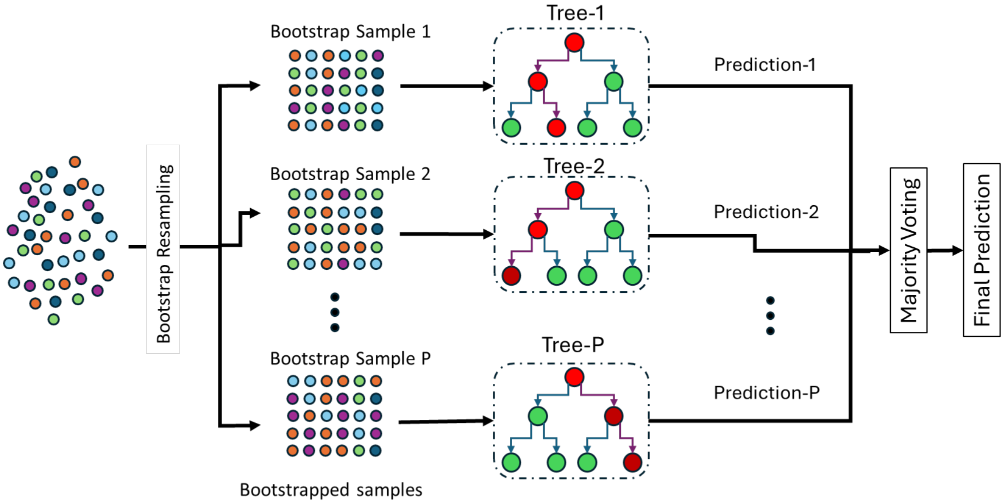


Figure 3. Workflow of the random forest.

2.5. Mutual information (MI)

Information theory enables the quantification of the linear and nonlinear relationships between variables [27, 28]. Information entropy, proposed by Shannon, is a measure of the uncertainty of a random variable. As the value of entropy increases, the uncertainty of the random variable also increases. Information entropy of a random variable $X = (x_1, x_2, \dots, x_N)$ is denoted as $H(X)$ and defined as shown in Equation (8).

$$H(X) = -\sum_{i=1}^{N} p(x_i)\, log\big(p(x_i)\big) \tag{8}$$

where $p(x_i)$ equals the probability of observing outcome x_i for variable X. The joint entropy of the two random variables $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_m)$ can be calculated as in Equation (9).

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) log(x_i, y_j) \tag{9}$$

where $p(x_i, y_j)$ are the probabilities of $n = i$ and $m = j$, respectively. If the value of variable Y is known, the conditional probability of X can be expressed mathematically as shown in Equation (10).

$$I(X;Y) = H(X) - H(X|Y) \tag{10}$$

Using the Equations (8) and (9), Equation (10) can be described as in Equation (11).

$$I(X;Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\frac{p(x_i|y_j)}{p(x_i)} \tag{11}$$

When $I(X;Y)$ is high, it means that X and Y are closely related to each other. If X and Y are independent, $I(X;Y)$ becomes 0, which indicates that knowledge of one variable does not give any information about the other.

3. Results and Discussion

This section presents experimental evaluations of benchmarked machine learning methods in obesity level prediction. Initially, the performance metrics utilized to evaluate the success of machine learning methods are described in detail. The performance of each method is then analyzed and visualized for method-specific hyperparameter ranges. The importance of the features was determined by the MI method and the effect of the features on machine learning was analyzed with the method with the best performing method among the compared methods.

3.1. Performance metrics

Performance metrics have been developed to assess the success of machine learning methods and to compare methods with each other. Confusion matrix, also known as contingency table, provides information about the correct and incorrect predictions of the machine learning model. The standard structure of a confusion matrix for a multi-class classification task is given in Table 3. In Table 3, the classes are labeled as $C_1, C_2, \dots, C_M$ and N_{ij} denotes the number of samples where the true class is C_i but is predicted as class C_j, where $i, j = 1,2, \dots, M$ and M is the number of classes.

Table 3. Confusion matrix for a classification task involving M classes.

		Predicted Class		
		C_1	$\cdots C_j \cdots$	C_M
Actual Class	C_1	N_{11}	N_{1j}	N_{1M}
	$\vdots$		$\vdots$	
	C_i	N_{i1}	$\cdots N_{ij} \cdots$	N_{iM}
	$\vdots$		$\vdots$	
	C_M	N_{M1}	N_{Mj}	N_{MM}

Various performance measures are generated from the confusion matrix. For imbalanced multi-class classification datasets, macro-averaged metrics are commonly employed. In the calculation of the macro-averaged metrics, the desired metric is computed for each class individually, and then the average is obtained by summing the results for all classes and dividing by the total number of classes. This approach assigns equal weight to each class, regardless of the sample size in each class. Widely adopted performance metrics for multi-class classification are accuracy, MAR, MAP, and macro-average F1-score.

Accuracy is described as the ratio of the total number of correctly classified samples in all classes to the number of samples in the dataset. It is defined as in Equation (12).

$$Accuracy = \frac{\sum_{i=1}^{M} N_{ii}}{\sum_{i=1}^{M} \sum_{j=1}^{M} N_{ij}} \tag{12}$$

MAR assesses the model's capability to accurately classify instances within each class. MAR is computed as in Equation (13).

$$Macro\ average\ recall\ (MAR) = \frac{1}{M} \sum_{i=1}^{M} \frac{N_{ii}}{N_{i.}} \tag{13}$$

where

$N_{i.} = \sum_{j=1}^{M} N_{ij} \quad \forall\, i \in \{1,2,..,M\}$

MAP evaluates the model's capacity to generate accurate correct predictions for every class. MAP is calculated as in Equation (14).

$$Macro\ average\ precision\ (MAP) = \frac{1}{M} \sum_{i=1}^{M} \frac{N_{ii}}{N_{.i}} \tag{14}$$

where
$N_{.i} = \sum_{j=1}^{M} N_{ji} \quad \forall\, i \in \{1,2,..,M\}$

The macro average F1-score considers both false positive and false negative predictions when computing the metric, enables a balanced evaluation of the model's overall performance. Macro average F1-score can be expressed mathematically as shown in Equation (15).

$$Macro\ average\ F1 - score = 2 \times \frac{MAP \times MAR}{MAP + MAR} \tag{15}$$

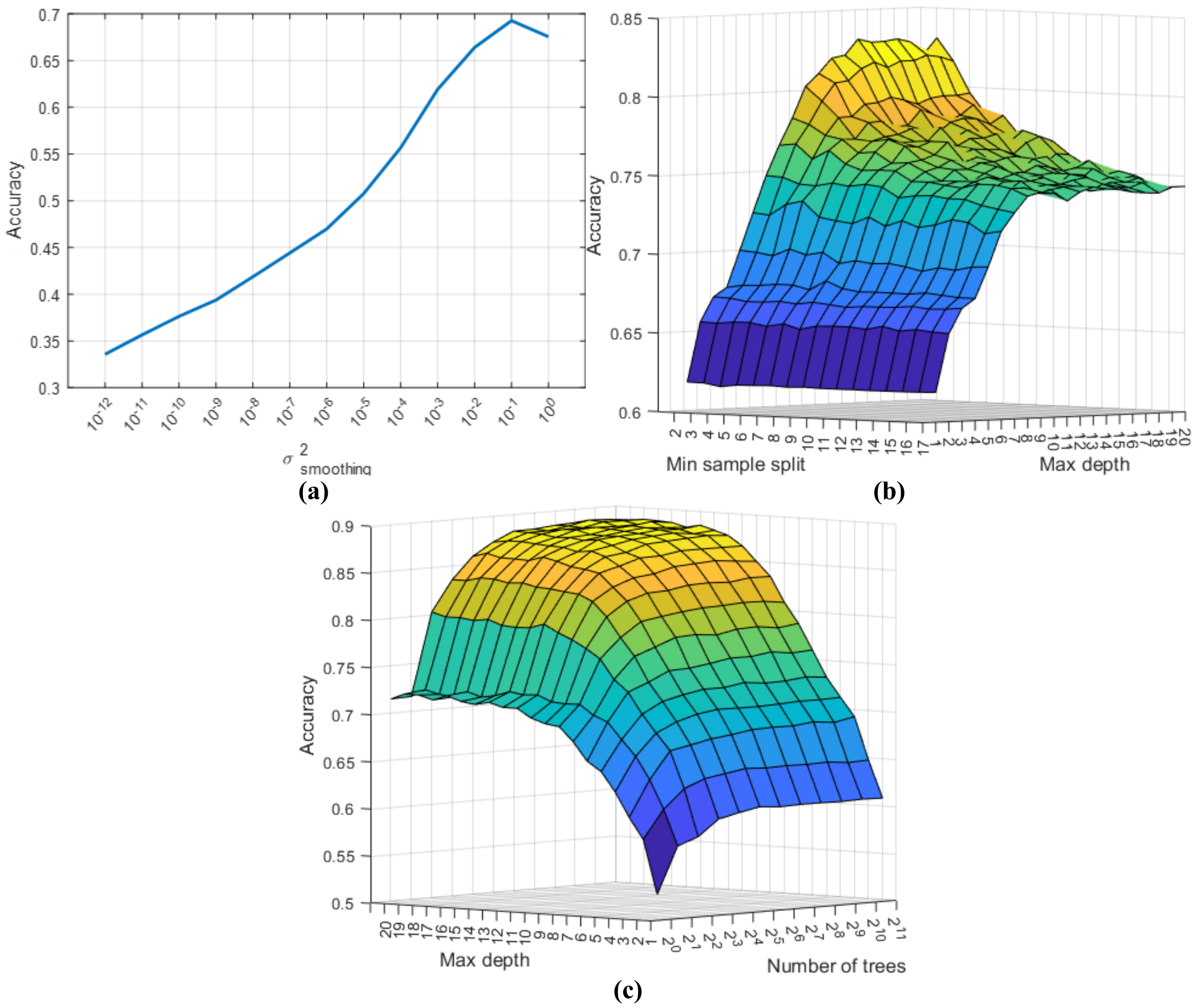
3.2. Experimental results

Obesity level was predicted by three different machine learning methods: NB, DT, and RF. These methods receive user-specified hyperparameters and the setting of these hyperparameters affects the performance of the methods. Therefore, in order to achieve the highest performance, method-specific hyperparameters need to be tuned to the dataset in each method. In the tuning of hyperparameters, determination is made using the brute-force grid search technique. The hyperparameters of each compared method, along with their corresponding ranges, are given in Table 4. To obtain more reliable performance estimation, k-fold cross validation is used. In the experiments, a k value of 10 was selected for the k-fold cross validation method, and the process was iterated 10 times. In the study, the average of 10 iterations obtained in the experimental results is given.

Table 4. Method specific hyperparameter range of the compared methods.

Method	Hyperparameter	Value
Naive Bayes	var_smoothing	$10^{-12}, 10^{-11}, \ldots, 10^{0}$
Decision Tree	Minimum samples split Maximum depth	2, 3, … ,17 1, 2, … , 20
Random Forest	Number of trees Maximum depth	$2^{0}, 2^{1}, \ldots, 2^{11}$ 1, 2, … , 20

The accuracy results of the method-specific hyperparameters of the compared methods are shown in Figure 4. The results of the NB method with the varying values of *var_smoothing* hyperparameter is shown in Figure 4(a). The accuracy of the methods increases as the value of *var_smoothing* approaches 10^{-1}; beyond that point, it starts to decrease. The highest accuracy is achieved at a *var_smoothing* value of 10^{-1}. The results of the DT method for the hyperparameter pair (*max_depth, number_of_trees*) are shown in Figure 4(b). The accuracy value increases with the increase of the *maximum_depth* hyperparameter. However, at low *max_depth* values (1 - 8), changes in the *min_samples_split* hyperparameter at the same *max_depth* values have little effect on the result. The highest values are obtained in the hyperparameter range where *max_depth* is high and *min_samples_split* is low. The performance of the RF algorithm concerning the *number_of_trees* and *max_depth* hyperparameters is illustrated in Figure 4(c). In the RF model, the highest accuracy values for the examined *max_depth* and *number_of_trees* hyperparameters were observed within the ranges 12–20 and 2^{6}-2^{11}, respectively. At values outside this range, the accuracy value begins to decrease gradually. The lowest accuracy is obtained when both *max_depth* and *number_of_trees* hyperparameters are at their minimum values.



(a)

(b)

(c)

Figure 4. Accuracy results of the machine learning methods a) NB, b) DT, c) RF.

3.3. Feature selection

With the development of technology, data collection has become very easy. In this way, a large number of attributes can be easily collected for a specific problem. Although it is generally thought that increasing the number of features will better define the problem and increase the performance of machine learning algorithms, this may not always be the case. A large number of features increase the complexity of the machine learning algorithms and the processing time. For this reason, determining the most prominent features in the collected data set and building the machine learning model according to these features will both reduce the required processing power and the complexity. In this study, the importance of the features in the dataset was determined using the MI method and they are ranked in descending order in Figure 5. As can be seen from Figure 5, age, vegetable consumption, and main meals are the most prominent features, while screen exposure, liquid consumption, and overweight/obese family are the least effective.

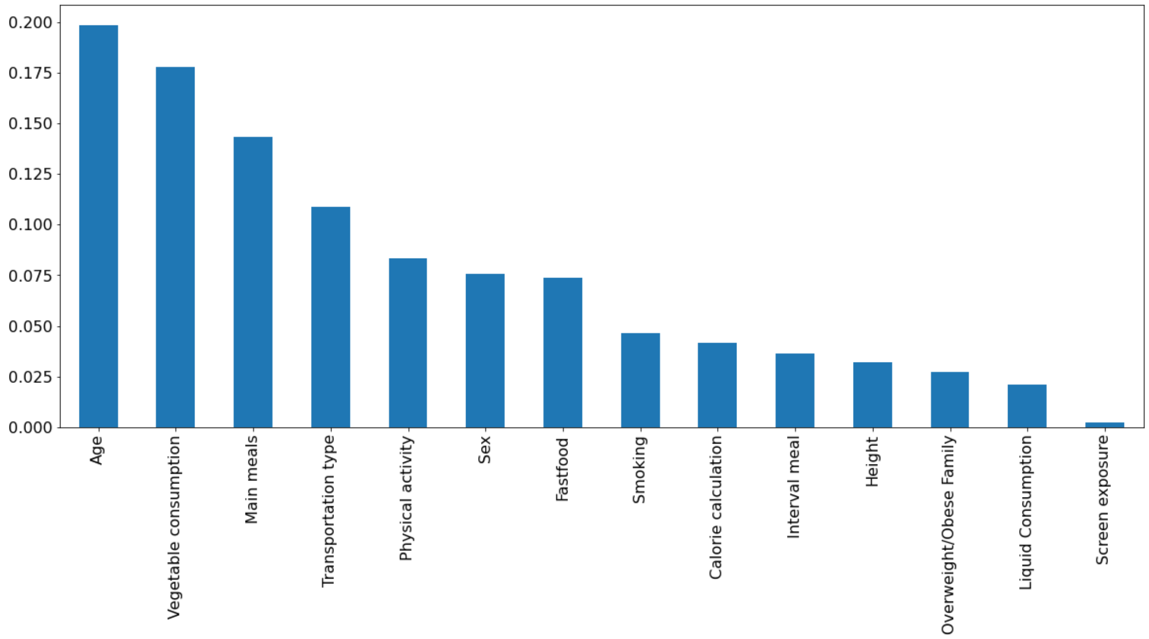
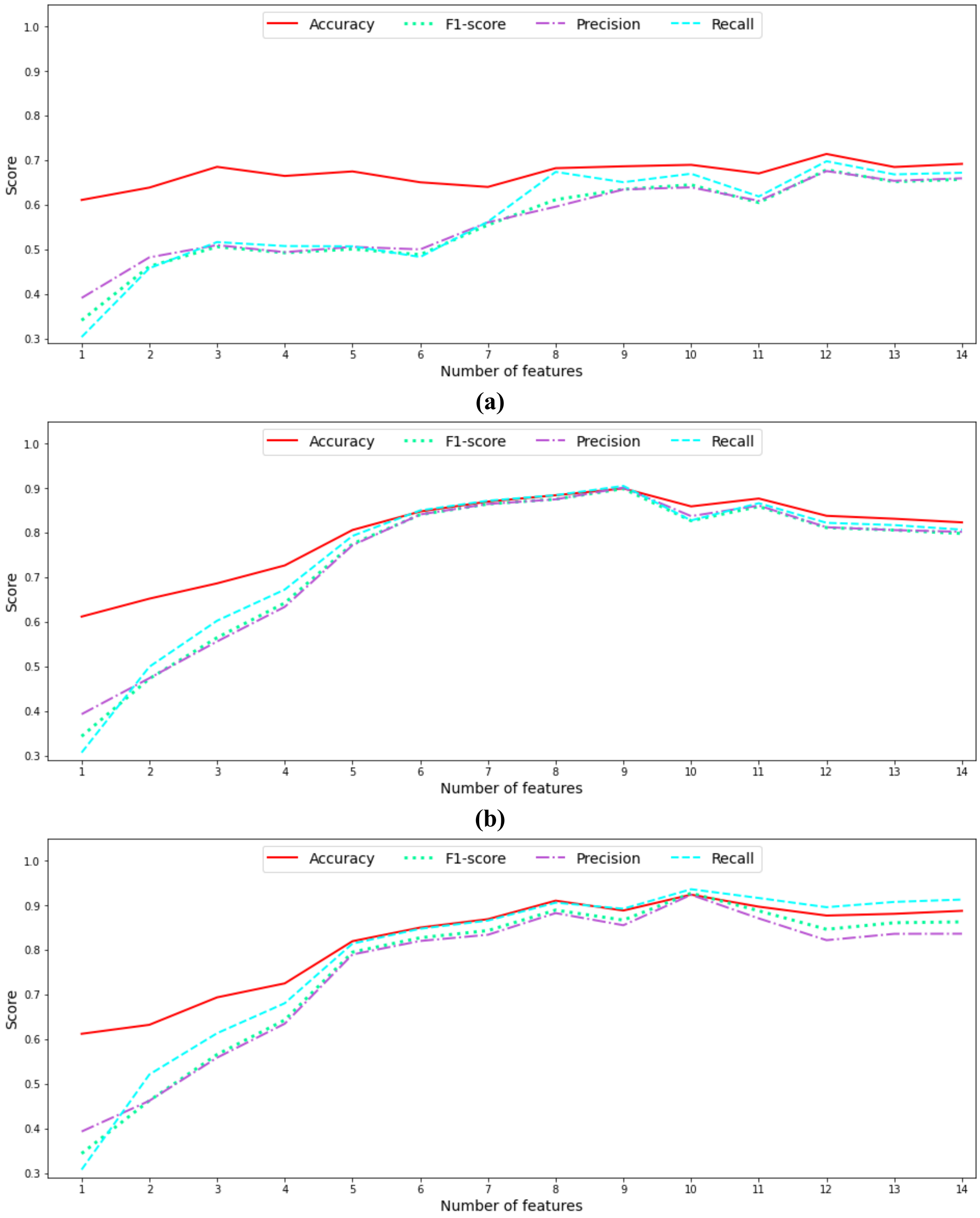


Figure 5. MI score of the features.

In order to observe the effect of the features in the dataset, the MI scores of the features were added one by one starting from the highest to the lowest, and data subsets were created. In the compared machine learning methods, models were created from these data subsets by using the hyperparameters that provided the best performance with all features in the data set. The results of these models were then evaluated with 4 different performance metrics to observe the effectiveness of each added feature. In this way, it will be possible to identify the subset of data with the highest performance using the minimum number of features. The experimental results obtained by the methods depending on the increasing number of features are shown in Figure 6.

In Figure 6(a), performance results of the NB method are presented. In NB, the highest results are obtained by using 12 features. Despite using 2 fewer features than the number of features in the entire dataset, higher results are obtained. Figure 6(b) shows the results of the DT method. The highest performance metric results for DT are obtained when the number of features is equal to 9. As can be seen in Figure 6(b), as the number of features increases from 1 to 9, the performance of the method increases by 47.03% in the accuracy performance metric. However, when comparing the results obtained with feature selection with the results obtained when all features are added, a performance improvement of 8.05% is obtained despite using 35.71% fewer features. Figure 6(c) demonstrates the performance results of the RF method. The highest results in the RF method were obtained by using the 10 features with the highest MI scores. After the highest values were obtained with 10 features, it was seen that increasing the number of features caused a slight decrease in performance.

Table 5 presents the performance results (accuracy, macro average F1-score, MAP, and MAR) of the compared methods with the determined hyperparameters, using both the entire set of features and a subset of features from the dataset. The hyperparameters of the methods were determined based on the values that achieved the highest results in the accuracy performance metric. In the remaining performance metrics, the results corresponding to these hyperparameters are given. In the table, the number of features for which the highest values were obtained by feature selection is given in the columns corresponding to the row named "Feature". When the NB method results in Table 5 are analyzed, it is seen that by using fewer features, a 3.03% increase in accuracy, a 3.34% increase in macro average F1-score, a 0.34% increase in MAP, and a 6.54% increase in MAR is achieved. The 4th column in Table 5 presents the results of the DT method. The highest percentage increase in performance with feature selection is observed for the DT method. The results of the RF method are given in the last column of Table 5. As in the other methods, higher results are obtained with feature selection in this method as well. Additionally, the highest performance results in all cases were achieved using the RF method with 10 features selected through the MI feature selection technique.



(a)

(b)

(c)

Figure 6. Performance results of the compared methods based on the increasing number of features **a)** NB, **b)** DT, **c)** RF

The only study in the literature using same data set was conducted by Köklü and Sulak [22]. In their study, the performances of three different machine learning methods, namely KNN, RF, and SVM, were compared. According to the results obtained, the highest accuracy was achieved with RF, similar to the presented study. However, the accuracy was improved by 1.25% in the current study. Furthermore, feature selection was performed with MI, which improved accuracy by 5.17% while reducing the number of features by 28.57%.

Table 5. Best performance results of the compared methods with/without feature selection.

Method	**Metric**	**Naive Bayes**	**Decision Tree**	**Random Forest**

Hyperparameter		var_smoothing=10^{-1}	maks_depth=14
			min_samp_split = 2
				maks_depth=19
				number_of_trees=2^9

Features		12	9	10
All Features	**Accuracy**	0.6927	0.8321	0.8892
	F1-score	0.6563	0.8041	0.8618
	MAP	0.6730	0.8090	0.8350
	MAR	0.6546	0.8121	0.9122
Feature Selection	**Accuracy**	0.7137	0.8991	0.9236
	F1-score	0.6782	0.8986	0.9269
	MAP	0.6753	0.9009	0.9232
	MAR	0.6974	0.9047	0.9358

To summarize, in all the compared machine learning methods, better performance was achieved using fewer features when feature selection was applied. The highest performance was achieved with the RF method both through features from the whole dataset and through feature selection. However, when comparing the performance results using all features in the dataset with those obtained through feature selection, the greatest improvement was observed in the DT method, with 8.05% in accuracy, 11.75% in macro average F1-score, 11.36% in MAP, and 11.40% in MAR. Although there is a 6.86% difference in accuracy performance metric between the DT and RF methods in the results of the experiments using all attributes, this difference decreased to 2.72% in the results after feature selection.

4. Conclusion

In this study, a detailed analysis of the performance of three different machine learning methods—NB, DT, and RF—was conducted to determine obesity levels. To achieve optimal performance, method-specific hyperparameter tuning was carried out using a brute-force grid search. Additionally, to enhance the reliability of the results, k-fold cross-validation (*k = 10*) was employed and repeated 10 times with the average results reported. According to the results, the RF method outperformed the others across four different performance metrics. Moreover, the most influential features for predicting obesity levels were identified using the MI method. Data subsets were then generated based on MI scores, and the performance of the methods, using the best hyperparameters identified from the entire dataset, was analyzed. While the DT method exhibited the greatest performance improvement, the RF method consistently achieved the highest overall performance. Considering both the smallest subset of features and the accuracy of predictions, RF was proven to be the most effective method for predicting obesity levels. When the results of two different experiments are examined together, by applying feature selection, a performance increase of 3.87% was achieved in the same method by using 28.57% fewer features.

The current study presented promising information for obesity determination. Thus, in the future, the presented study is open to improvements in the following aspects. Different deterministic features may be added to enhance the performance while removing the less relevant ones. Furthermore, explainable machine learning methods can be implemented to learn the decisions behind the constructed machine learning models. Hence, the trustworthiness of the model predictions can be enhanced.

References

[1] World Obesity Federation. “World Obesity Atlas 2023.” Available: https://data.worldobesity.org/publications/?cat=19

[2] Włodarczyk M, Nowicka G. Obesity, DNA damage, and development of obesity-related diseases. Int J Mol Sci 2019; 20(5): 1146.

[3] Mohajan D, Mohajan HK. Obesity and its related diseases: a new escalating alarming in global health. J Innov Med Res 2023; 2(3): 12-23.

[4] Göktaş ÖF, Çankaya İ, Ermeydan EŞ. Determination of the Optimum Test Conditions for Measurement of Glucose Level in Liquids. TJST 2024; 19(1): 45-53.

[5] Okunogbe A, Nugent R, Spencer G, Powis J, Ralston J, Wilding J. Economic impacts of overweight and obesity: current and future estimates for 161 countries. BMJ Glob Health 2022; 7(9): e009773.

[6] World Health Organization. (2024). Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[7] Nuttall FQ. Body mass index: obesity, BMI, and health: a critical review. Nutr Today 2015; 50(3): 117-128.

[8] De Koning L, Merchant AT, Pogue J, Anand SS. Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: meta-regression analysis of prospective studies. Eur Heart J 2007; 28(7): 850-856.

[9] Degirmenci A, Karal O. iMCOD: Incremental multi-class outlier detection model in data streams. Knowledge-Based Syst 2022; 258: 109950.

[10] Degirmenci A, Karal O. Efficient density and cluster based incremental outlier detection in data streams. Inf Sci 2022; 607: 901-920.

[11] Özbay FA, Özbay E. An NCA-based hybrid cnn model for classification of Alzheimer's disease on grad-cam-enhanced brain MRI images. TJST 2023; 18(1): 139-155.

[12] Degirmenci A. Performance comparison of kNN, random forest and SVM in the prediction of cervical cancer from behavioral risk. Int J Innov Sci Res Technol 2022; 7(10): 71-79.

[13] Peeyada P, Cholamjiak W. A new projection algorithm for variational inclusion problems and its application to cervical cancer disease prediction. J Comput Appl Math 2024: 441, 115702.

[14] Goktas OF, Demiray E, Degirmenci A, Cankaya I. Real time non-invasive monitoring of glucose and nitrogen sources with a novel window sliding based algorithm. Eng Sci Technol Int J 2024; 58: 101845.

[15] Cheng ER, Steinhardt R, Ben Miled Z. Predicting childhood obesity using machine learning: Practical considerations. BioMedInformatics 2022; 2(1): 184-203

[16] Solomon DD, Khan S, Garg S, Gupta G, Almjally A, Alabduallah BI, Alsagri HS, Ibrahim MM, Abdallah AMA. Hybrid Majority Voting: Prediction and Classification Model for Obesity. Diagnostics 2023; 13(15): 2610.

[17] Kaur R, Kumar R, Gupta M. Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence. Endocrine 2022; 78(3): 458-469.

[18] Wang Q, Yang M, Pang B, Xue M, Zhang Y, Zhang Z, Niu W. Predicting risk of overweight or obesity in Chinese preschool-aged children using artificial intelligence techniques. Endocrine 2022; 77(1): 63-72.

[19] Liu W, Fang X, Zhou Y, Dou L, Dou T. Machine learning-based investigation of the relationship between gut microbiome and obesity status. Microbes Infect 2022; 24(2): 104892.

[20] Wong JE, Yamaguchi M, Nishi N, Araki M, Wee LH. Predicting overweight and obesity status among Malaysian working adults with machine learning or logistic regression: retrospective comparison study. JMIR Format Res 2022; 6(12): e40404.

[21] Calderón-Díaz M, Serey-Castillo LJ, Vallejos-Cuevas EA, Espinoza A, Salas R, Macías-Jiménez MA. Detection of variables for the diagnosis of overweight and obesity in young Chileans using machine learning techniques. Procedia Comput Sci 2023; 220: 978-983.

[22] Koklu N, Sulak SA. Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities. Sinop Uni J Nat Sci 2024; 9(1): 217-239.

[23] Koklu N, Sulak SA. Obesity Dataset. Kaggle. https://www.kaggle.com/datasets/suleymansulak/obesity-dataset: 2024.

[24] Kim T, Lee JS. Maximizing AUC to learn weighted naive Bayes for imbalanced data classification. Expert Syst Appl 2023; 217: 119564.

[25] Tokgöz N, Değirmenci A, Karal Ö. Machine Learning-Based Classification of Turkish Music for Mood-Driven Selection. J Adv Res Nat Appl Sci 2024; 10(2): 312-328.

[26] Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens 2012; 67: 93-104.

[27] Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948; 27(3): 379-423.

[28] Cover TM. Elements of information theory. John Wiley & Sons.