# Age-specific risk factors for the prediction of obesity using a machine learning approach

**Junhwi Jeon**
Kyung Hee University
**Sunmi Lee**
Kyung Hee University
**Chunyoung Oh** ( corresponding.cyoh@jnu.ac.kr )
Chonnam National University

**Article**

**Keywords:**

**Posted Date:** May 5th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1515734/v1

# Age-specific risk factors for the prediction of obesity using a machine learning approach

**Junhwi Jeon**[1,+], **Sunmi Lee**[1,+], **and Chunyoung Oh**[2,*]

[1]Kyung Hee University, Department of Applied Mathematics, Yongin, 17104, Korea
[2]Chonnam National University, Department of Mathematics Education, Gwangju, 61186, Korea
[*]corresponding.cyoh@jnu.ac.kr
[+]These authors contributed equally to this work

## ABSTRACT

Machine Learning is a powerful tool to discover hidden features in various data driven research fields. Obesity involves extremely complex factors, such as biological, physiological, psychological, and environmental factors. A machine learning framework can provide a successful approach to revealing essential risk factors of the complex obesity phenomenon. Over the last two decades, the obesity population (BMI of above 23) in Korea has been rapidly growing. In this work, we assess obesity prediction by utilizing eight Machine Learning algorithms, and identify risk factors of obesity based on the Korea National Health and Nutrition Examination Survey (KNHANES) data 2016-2019. We explore age-specific and gender-specific risk factors of obesity for adults (19-79 years old). Our findings show that the risk factors for obesity are sensitive to age and gender under different Machine Learning algorithms. Both male and female 19-39 age groups show the highest performance of over 70% accuracy and ROC while the 60-79 group shows around 65% accuracy and ROC. Both male and female 40-59 age groups achieved the highest performance of over 70% in ROC, but the female achieved lower 70% in accuracy. Our results highlight that the top four significant features in all age gender groups for prediciting obesity are Triglyceride, ALT(SGPT), Glycated hemoglobin, and urine acid. For the accuracy ratio of the classifiers and age groups, there is no big difference in accuracy when the number of features is more than six, except the accuracy ratio decreased in the female 19-39 age group.

## 1 Introduction

Obesity prevalence has become one of the most prominent issues in global public health. The causes of obesity were affected in several categories, of physiology, individual psychology, food production, food consumption, physiology, individual physical activity, genetic and cultural influence, and physical activity environment etc.[1,2]. With the number of obese people doubling in two decades (from 1.3 million people obese globally in 1980 to double that in 2008), unhealthy habits, unhealthy diet, intake of high saturated fat, and discretionary foods, and physical inactivity are the major pillars of "obesity and overweight"[3]. Some disorders are caused by mutation in a single gene, while many other are much more complex. Some medical problems, like obesity, do not have a single genetic cause, and they are likely to be associated with the effects of multiple genes in combination with environmental factors and lifestyle. The medical problems caused by obesity increases the risk of other diseases and health problems, such as heart disease, diabetes, high blood pressure, and certain cancers. Moreover, obesity can diminish the overall quality of a person's life.

Obesity is responsible for a large fraction of costs, to both the health care system, and to society at large. Diabetes, cancer, cerebrovascular disease, hypertensive disease, and authorities were the diseases related to obesity, which resulted in socioeconomic costs of about KRW 1.36 trillion in long-term socioeconomic costs associated with

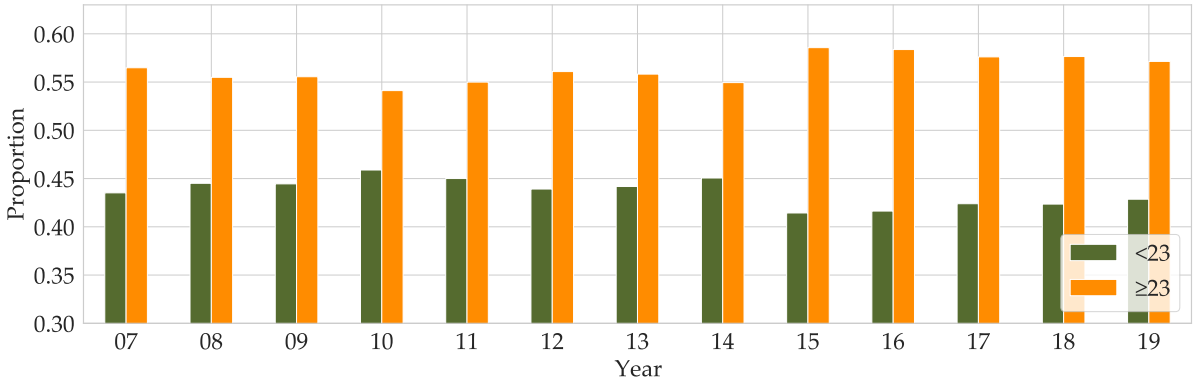


**Figure 1.** Annual proportions of BMI below and above 23 are shown from 2007 to 2019. (Orange for BMI above 23 and green for BMI below 23).

adolescent obesity[4]. As of 2016, social costs, such as medical and nursing expenses due to obesity, were estimated at 11.5 trillion (KRW) per year in Korea[5]. The review for Europe which encompassed both direct and indirect costs estimated obesity-related costs to range (0.09 to 0.61%) of total annual gross domestic income in Western European countries[6,7]. Moreover, social and economic factors are linked to obesity. If the costs of illness attributable to obesity could be minimized, monetary resources within the national health care systems and economies could be re-allocated toward other ends.

Until the 1970s, obesity was defined by reference to an 'ideal body weight', derived from actuarial tables compiled by the life insurance industry. The Body Mass Index (BMI) in adults is defined as the ratio of body mass in kilograms to the square of the individual's height in meters. In the 1980s, the ideal body weight approach was replaced by BMI($kg/m^2$), and the commonly used cutoffs for normal weight (BMI:18.5 $\sim$ 22.9), overweight (BMI: 23 $\sim$ 24.9) and obesity (25 $\leq$ BMI), for both men and women, were adopted to define obesity in adults according to the Asia-Pacific guidelines[8,9]. BMI and waist circumference are known risk factors for obesity criteria. Even though BMI leads to confusion and misinformation, BMI has been the most commonly used measure of adiposity in epidemiological research. We used BMI as our obesity criterion. The Korea National Health and Nutrition Examination Survey (KNHANES) has been conducted since 1998, and the BMI factor was included from 2007 in the survey. The KNHANES is one of the principal sources for investigating obesity in the population with vast amounts of variables and data related to obese diseases. According to the BMI data of KNHANES, in 2007, overweight adults were 23% and obese were 32%; while in 2019, the overweight adults were 23% and obese were 34%. Over 12 years, the obese increased by 2%, while there was no increase in overweight[10] (see Fig. 1). Overall, the proportion of the sum of overweight and obesity is much higher than that of a BMI of less than 23.

Currently, the use of machine learning models for disease classification has been increasing rapidly, both because of the significant amount of data available that is being generated by healthcare devices and systems, and the magnitude of computational resources available for data calculation and processing[11–13]. Obesity researchers and health care professionals have access to a wealth of data. Importantly, this immense volume of data is utilized to train models, and facilitates the use of expert systems, machine learning techniques, and classification techniques for finding trends and patterns in the evaluation and classification of several diseases[14]. Machine learning techniques applied to large survey data sets may provide a meaningful data-driven approach to categorize patients for population health management, which is part of the critical factors for obesity. Machine learning algorithms can be applied to assess the factors leading up to the prevalence of obesity. Our approach to predict obesity was to use machine learning on the KNHANES datasets of South Korea to estimate obesity.

As obesity increases, so does the risk for a variety of diseases; identifying suspected clinical findings among the top-ranked factors is important to preventing obesity in people. In this paper, we aimed to identify the metabolic factors affecting BMI or obesity in KNHANES datasets, and to predict obesity utilizing various machine learning algorithms. First, we carried out correlation analysis among the numerous risk factors for obesity. Second, we split the KNHANES dataset into six groups according to gender and three age groups. Third, we employed eight machine learning algorithms: Multi-Layer Perceptron (MLP), Random Forest (RFC), Gradient Boosting (GBM), support vector machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), Light Gradient Boosting (LGBM), and Extreme Gradient boosting (XGB). In addition, we measured the performance of machine learning classification algorithms in terms of certain performance metrics (specificity, accuracy, ROC, precision, recall, F1-score, etc.). Furthermore, we identified the key top-ten age-specific and gender-specific risk factors through feature importance methods (Random forest and Sharp value). Based on the top-ten risk factors, we determined the impact of these factors on the performance (accuracy) of the machine learning algorithms.

## 2 Literature review

Numerous studies have used different methods of machine learning to predict obesity using diverse factors as features. This section of the paper deals with all of the past and present work that predicted the risk of adult obesity using machine learning algorithms(see Table 1).

Xiaolu Cheng et al.[15] classified using data from NHANES. They predicted whether an individual was overweight or obese based on physical activity levels. Respondents wore ActiGraph AM-7164 devices to record their physical activity. Eleven algorithms—logistic regression, naïve Bayes, Radial Basis Function, local k-nearest neighbors, classification via regression, random subspace, decision table, multi-objective evolutionary fuzzy classifier, random tree, J48, and multi-layer perceptron—were implemented and evaluated, and compared with traditional logistic regression model estimates. The mean percentages of overall accuracy, sensitivity, and specificity of the eleven classifiers were (62.37, 70.89, and 49.70)%, respectively. Ferdowsy et al.[17] applied nine machine learning algorithms for k-nearest neighbor, random forest, logistic regression, multilayer perceptron, support vector machine, Naïve Bayes, adaptive boosting (ADAB), decision tree, and gradient boosting (GBM) classifier. They used the user's daily activities, food routines, height, and weight. They obtained the best performance from the model, the logistic regression algorithm, which achieved the highest accuracy of 97.09%, compared to the other classifiers. The gradient boosting algorithm gave the accuracy of 64.08%, as well as the lowest metric values.

Chatterjee et. al[18] performed a regression analysis to visualize the trend of change in age, tobacco consumption, sweet beverages, economic condition, fast food, sleeping pattern, diet, blood pressure, blood glucose, lipid profile, adiposity, exercise, and family history in relation to obesity/overweight change diabetes type II in the sample population, excluding genetic factors and pregnancy. They used support vector machine with linear and radial basis function kernel, Naïve Bayes, Decision Tree, Random Forest, and 'KNN' models, but 'SVM' with linear kernel provided the best classification, with accuracy of 0.95. Dunstan et al.[20] used three non-linear machine learning algorithms—SVM, Random Forest, and Extreme Gradient Boosting to predict obesity incidence at the country level, based on countrywide sales of a small subset of food and beverage classes. The study predicted that baked goods and flours, followed by cheese and sweet carbonated drinks, were the most pertinent food categories to predict obesity. Jindal et al.[21] performed ensemble machine learning approaches for obesity prediction based on the key determinants of age, height, weight, and 'BMI'. The ensemble model utilized Random Forest, generalized linear model, and partial least square, with a prediction accuracy of 89.68%.

To predict the future risk of developing complex diseases such as obesity, based on BMI status and SNP profile, Abdulaimma et al.[22] introduced a genetic profile predictive study using machine learning algorithms, and used the publicly available participants' profiles, genetic variants, or Single Nucleotide Polymorphisms (SNPs). They found 13 SNPs. Seven machine learning algorithms for the prediction of obesity were used on the 13 SNPs, and support vector machine generated the highest area under the curve value of 90.5%. Grabner at al.[23] performed a study on the National Health and Nutrition Examination Survey (NHANES), National Health Interview Survey (NHIS), and

**Table 1.** Machine learning models and the risk factors related to adult obesity

| Author | Type of model | Risk Factors |
|---|---|---|
| Xiaolu Cheng et al,[15] | LR, Naïve Bayes, Radial Basis Function, Local KNN, Random subspace, J48, Decision table Random tree, MLP | physical activity in NHANES |
| Sri Astuti Thamrin et. al.[16] | LR, Naïve Bayes, CART | location, marital status, age group,education, work category, sugary foods,sweet drinks,instant foods, energy drinks, salty foods , fatty/oily foods, grilled foods, preserved foods, smoking, seasoning powders,soft/carbonated drinks, alcoholic drinks, mental-emotional disorders, diagnosed hypertension, physical activity, fruit and vegetables consumptions, RISKESDAS |
| Ferdowsy et. al[17] | KNN, RF, LR, GB, MLP, SVM, Naïve Bayes, ADAB, DT, GBM | Daily activities, Food routines, Height, Weight |
| Chatterjee et. al.[18] | SVM, Naïve Bayes, DT, RF, KNN | Age, Tobacco consumption, Fast food, Sleeping pattern, diet, Blood pressure, Blood glucose, Lipid profile, Adiposity, Exercise, Family history, Sweet beverages, Economic condition |
| Singh et. al.[19] | Multi LR, ANN | Early BMI |
| Dunstan et al.[20] | SVM, RF, XGB | Sales of food and Beverage classes |
| Jindal et. al.[21] | Ensemble utilized RF | Age, Height, Weight, BMI |
| Montañez et. al.[22] | GBM, GLMNET, CART, KNN, SVMRadial, RF, NNET | Age, Gender, Genetic variants or single Nucleotide Polymorphisms (13-SNPs) |

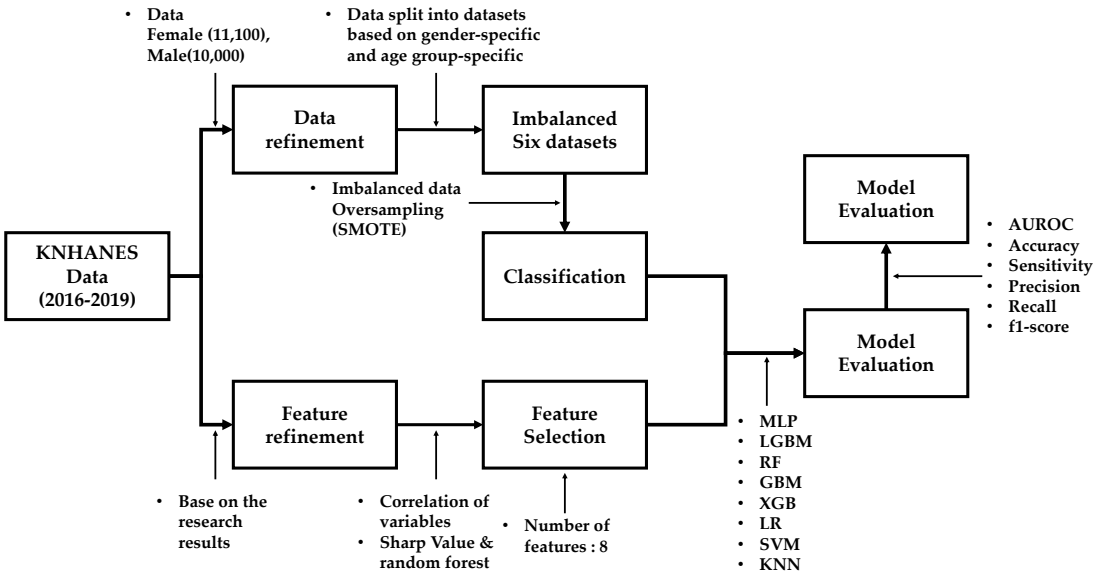


**Figure 2.** Outlines for the obesity prediction.

Behavioral Risk Factor Surveillance System (BRFSS) datasets from the 1970s to 2008 to analyze the trend of BMI in the USA over time, and across race, gender, socioeconomic background, and status. SES–BMI gradients were observed to be steadily more significant for women than for men. The model outperformed the traditional logistic regression model in terms of predictive power.

Thamrin et al.[16] used the three models of logistic regression, classification and Regression Trees (CART), and Naïve Bayes to identify the presence of obesity using the publicly available health data, RISKESDAS. Location, marital status, age group, education, sweet drinks, fatty/oily foods, grilled foods, preserved foods, seasoning powders, soft/carbonated drinks, alcoholic drinks, mental/emotional disorders, diagnosed hypertension, physical activity, smoking, and fruit and vegetable consumptions were significant in predicting obesity status in adults. Safaei et al.[1] reviewed the obesity literature published from 2010 to 2020. They introduced and reviewed machine learning techniques for obesity prediction that used a wide variety of features as risk variables.

According to the works of the review literature on obesity, although obesity has been predicted using different variables[1], the predictions of obesity for most of the variables utilized here were not used. In most studies of the prediction of obesity, the data used by the studies focused on certain aspects of the participants' lifestyles or behavior as features, however, our research utilized participants' metabolic factors in the KNHANES. To the best of our knowledge, there is no research that applied machine learning algorithms while utilizing only metabolic factors in the KNHANES data. Therefore, we were motivated to employ machine learning techniques to obtain satisfactory results of obesity prediction in a wide variety of situations. Figure 2 presents a flowchart of our approach to obesity prediction using machine learning algorithms.

## 3 Descriptions of data and features

### 3.1 Data sources

Our datasets are prepared and published through the Korea National Health and Nutrition Examination Survey (Korea Division of Health and Nutrition Survey and Analysis, KDCA), to provide full access. The Korea National

Health and Nutrition Examination Survey (KNHANES) is a national program that is designed to assess the health and nutritional status of adults and children in Korea. Since 1998, the KNHANES has collected data obtained by direct physical examination, clinical and laboratory tests, personal interviews, and related measurement procedures. The KNHANES was conducted on a triennial basis from 1998 to 2005. In 2007, the survey became a continuous, annual survey program conducted by the KDCA, and varieties of health measurements were added to the basic design to meet emerging data needs[24]. In 2007, the BMI factor was added, and uric acid factor has been included as a survey variable since 2016[10].

Two stages of stratified clustering−consisting of primary sampling units and households−were applied to the data collected from the population and housing census in Korea. In general, blood and urine samples are collected from participants aged 10 years and over[25]. The extent of examination differs, depending on the age of the participant, but targeted individuals start at the age of 1[10]. These data are used to estimate the prevalence of chronic disease in the total population, or monitor trends in the prevalence and risk behaviors. Approximately 10,000 persons are sampled in total in all 192 variables for the primary sample units per year[24]. These data are composed of demographic variables, health questionnaires, medical examination, and a nutritional survey[10,24]. The data in KNHANES consists of categorical data, numeric data, and text data, and we refined the numeric data as obesity risk factors. The dataset can be accessed at the Korea National Health and Nutrition Examination Survey. knhanes.kdca.go.kr/knhanes/eng/index.do.

### 3.2 Oversampling

In machine learning applications, data preprocessing plays a significant role in achieving better performance and accurate results. Our data were from 19 to 79 years old for the 2016-2019 years out of the KNHANES. We conducted data removing by excluding all records with incomplete or missing values for the variable/feature BMI, a core feature used to categorize obesity status. After cleaning missing values, the participants were about 21,100 for males and females for the top features selected following feature selection. After we split the dataset into two parts by gender, we divided both male and female into three age groups (19-39, 40-59, and 60-79), respectively. Therefore, our datasets consist of six datasets, as shown in Table S1 of the Supplementary Information. Each dataset used the training part to train the model (75%), and the testing part to test the model (25%).

As overweight and obese categories are at risk, instances belonging to these two classes have been combined, and have been labeled as 'BMI'. Although this slightly reduces the imbalance, the majority of the algorithms are only able to classify the majority class with a high degree of accuracy. One of the factors that deteriorates the accuracy of the test dataset in AI classification modeling is class imbalance. If the model is trained in a state where the number of each class is significantly different from the data, the predictive model tends to be biased toward a specific class, or makes it difficult to evaluate properly, acting as a factor of performance degradation. Conducting data analysis in a highly imbalanced data set is not trivial, and often leads to low sensitivity results being obtained[14]. Our dataset consists of six groups that are composed of gender and age groups. Table S1 of the Supplementary Information (SI) illustrates the majority of the instances belonging to the obesity category, which makes the datasets imbalanced. The Synthetic Minority Over-sampling Technique (SMOTE) is one of the representative oversampling techniques. It is a method to create a new sample, and add it to the data by using the k-NN($k \geq 2$) algorithm from a sample of a class with a small number of data[26]. In this study, the SMOTE technique with oversampling was used, which resulted in two new datasets of 50% and 50%. Since SMOTE generates data based on an algorithm, the possibility of overfitting is less than that of a simple random method.

### 3.3 Feature selection

Feature selection is used to identify the optimal set of the most important features that are capable of improving the performance of the classifiers. Obesity is caused by a number of factors, and a great many studies have identified various factors that cause obesity. However, obesity prediction in all factors in the KNHANES with models did not show reliable performance. We chose risk factors according to the following procedure. We narrowed down to the

**Table 2.** Metabolic factors causing BMI or obesity based on previous studies.

| Feature Name | References | Feature Name | References | Feature Name | References |
|---|---|---|---|---|---|
| BMI | | Red blood cell (RBC) | 28 | Systolic(diastolic) blood pressure (sdp)(dbp) | 29 |
| Hemoglobin (Hb) | 30 | ALT(SGPT) | 31,32 | Triglyceride (TG) | 33,34 |
| Fasting serum glucose (glu) | 35 | White blood cell (WBC) | 36 | Glycated hemoglobin (HbA1c) | 37 |
| Serum creatinine (cre) | 38 | Uric acid (UA) | 39,40 | Platelet (Bplt) | 41,42 |
| Ast(SGOT) | 43 | Cholestol (chol) | 44 | Hematocirit (Hct) | 28 |

top-15 factors out of the KNHANES contributing to BMI or obesity based on numerous previous research, and used factors obtained in the previous study[27](see Table 2). Those key factors were known as variables that contributed to increase BMI or caused obesity. Table 2 shows essential influential factors that determine adult overweightness or obesity.

Next, we carried out correlation analysis, and used predetermined factors obtained in the previous study[27] and Table 2. Figure 3 depicts the correlation between the metabolic features selected based on the KNHANES, and illustrates the correlation coefficient between the selected variables and the target feature, BMI, and the remaining independent factors. Figure 3 also identifies multicollinearity between independent factors. The top-10 features are included in Table 2 and Figure 3. We conducted feature importance analysis through a tree model, such as a random forest. However, due to the nature of decision trees that branch using Gini impurity, there is a tendency to overestimate the importance of continuous variables or variables with many categories that have many chances to be the basis for node branching. We additionally used SHAP feature importance. Shapley Additive explanation (SHAP) values have been proposed for SHAP values as a unified measure of feature importance[45]. SHAP feature importance defines feature importance as a mean absolute SHAP value. Each time a feature is added, the contribution is calculated and averaged. When used for tree-based models, SHAP has the great advantage of being able to calculate Shapley values relatively quickly. The selected features made up slightly different features according to the age group and gender. Therefore, we utilized it to identify the principal features in model prediction. we used the mean of the Random Forest and mean absolute Shap value, and selected ten factors for each of the six age-specific groups.

## 4 Machine learning algorithms

The feature vectors selected in the above way are fed to several classifiers for training, and then testing. We utilized eight classifiers, namely Multi-Layer Perceptron (MLP), Random Forest (RF), Gradient Boosting (GBM), Light Gradient Boosting (LGBM), Extreme Gradient Boosting (XGB), support vector machine (SVM), logistic regression (LR), and K-Nearest Neighbor (KNN). In the arena of performance analysis, accuracy cannot be claimed as a rigorous metric for the measurement of the actual performance of a classifier, because it may not be well fitted to estimating classification patterns obtained from an imbalanced set of data. Eight algorithms have been used for our research, and they each have certain parameters. These parameters have different values, which vary from one another. These parameter values are used to train the model, and they are discussed in Table 3.

The classification problem is only binary designed, because our data will only consider a model for binary classification based on BMI. A Multi-Layer Perceptron is made up of a large number of neurons. In general, it consists of three layers: an input layer, a hidden layer, and an output layer, while an activation function exists between each layer, giving non-linearity. In supervised learning, since inputs and outputs are given, only the weights

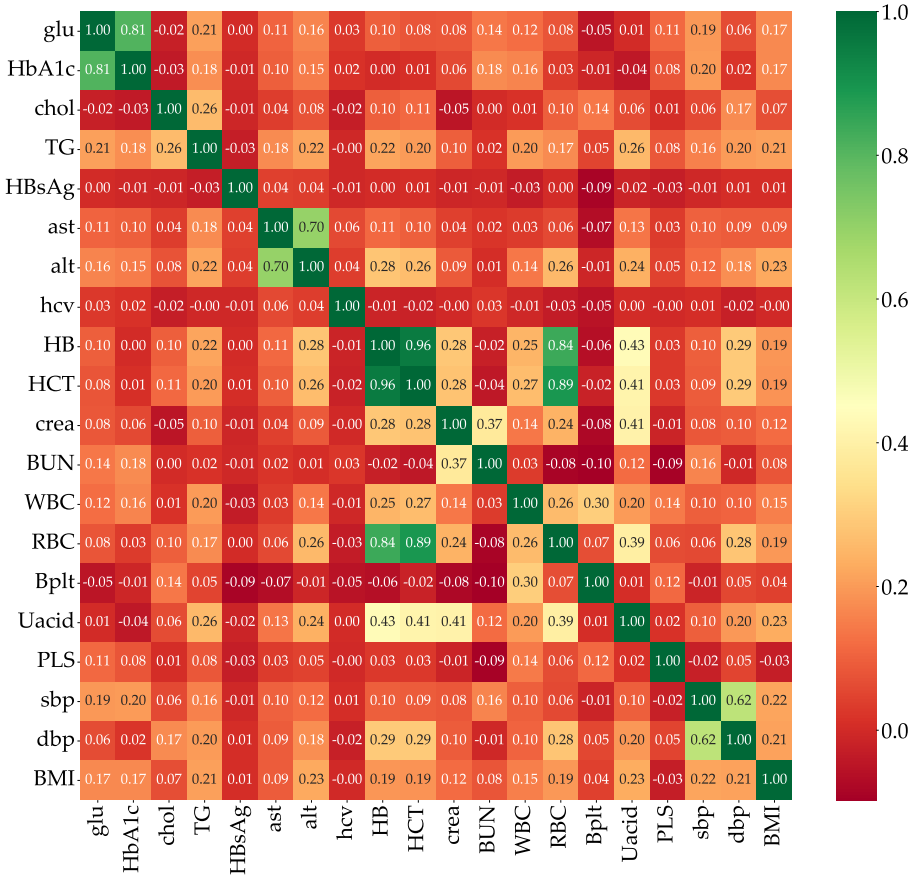


**Figure 3.** Correlation between BMI and 19 selected metabolic features given in KNHANES.

**Table 3.** Detailed specifications for the eight algorithms.

| Algorithms | Specification of algorithms |
|---|---|
| MLP | Alpha = [0.0001, 0.001, 0.01, 0.1] |
| | Activation : relu |
| | Hidden layer size [100 100] |
| GBM | Number of tree = [3, 10, 30, 40, 50, 70, 100] |
| | Maximum depth = [2, 3, 5, 10, 30] |
| RFC | Number of tree = [3, 10, 20, 30, 40, 50, 70, 100] |
| | Maximum depth = [2, 3, 5, 10, 30, 50 ] |
| KNN | Number of neighbor = [3, 5] |
| | Power parameter, p = 2 |
| | Metric : Minkowski |
| | Distances metric: Minkowski distance= $(\sum_{i=1}^{k}(\lvert x_i - y_i \rvert)^q)^{1/q}$ |
| SVM | C = [0.1, 1, 5, 10, 30, 50] |
| | Gamma = [0.5, 1, 10, 30, 50] |
| | Kernel: radial basis function |
| LR | Maximum number of iterations = 1000 |
| | Penalty = l2 |
| LGBM | Maximum depth =[2, 30, 50] |
| | number of leaves = [3, 70, 100] |
| XGB | Maximum depth =[2, 10, 30, 50] |
| | gamma = [0, 0.1, 0.2] |
| | number of estimators = [10, 20, 30, 40, 50, 70, 100] |

are updated. MLP utilizes back propagation for training, which is a supervised learning technique. Support vector machine (SVM) is mainly used for classification and regression analysis, and is a model that defines decision boundaries for classification. SVM builds a maximum margin separator, which is used to make decision boundaries with the largest possible distance. The classification algorithm of logistic regression is used to assign observations to an individual set of classes.

When new data is input, K-Nearest Neighbor (KNN) selects the k pieces of data closest to the new data by comparing existing data (neighbors) around the new data. Here, the most frequent labels in k data are assigned to new data. Logistic regression (LR) predicts the probability that data will fall into a certain category as a value between 0 and 1, and classifies it as belonging to a more likely category based on that probability. Logistic regression can be viewed as a classification technique, because the dependent variable is intended for categorical data. In particular, LR is mainly used when the dependent variable is a binomial problem. Random forest (RF) is a kind of ensemble learning method that integrates predictions of multiple base models, and outputs classification results from multiple decision trees constructed during the training process. So, Random forest chooses a voting method that collects classification results from a number of decision trees configured through training to obtain a conclusion. Random forest can quickly build a model even when the size of the data is huge, and the decision tree, which is the base model, has the advantage that it does not require a premise for data distribution.

Gradient Boosting (GBM) is a predictive model that can perform regression analysis or classification analysis, and is an algorithm belonging to the boosting family among ensemble methodologies. Boosting is the process of combining weak classifiers to create a strong classifier. Gradient boosting machines usually use a decision tree as the base model, starting with a single leaf, not a stump or tree. Light Gradient Boosting (LGBM) is a gradient learning framework based on decision tree and the idea of boosting[46], and is a relatively new model. It uses histogram-based algorithms to speed up the training process. GBM adds a maximum depth limit to the top of the leaf to prevent overfitting, while ensuring high efficiency. The Extreme Gradient Boosting (XGB) algorithm is a variant of gradient boosting based on decision trees. It provides parallel tree boosting, and is the leading machine learning algorithm for regression, classification, and ranking problems. Boosting is a common ensemble learning technique that combines sequentially weak learners to produce a powerful final learner. Each base learning algorithm learns from its previous base learner, and reduces its error. To predict the final output, the XGBoost algorithm combines the weights of the leaves from all trees.

For each of the six datasets, we measured the performance of a classifier based on the confusion matrix. It can be easily noticed that the 2-class is a $2 \times 2$ matrix. A binary states the number of true positives (TPs), false negatives (FNs), false positives (FPs), and true negatives (TNs) for a 2-class problem. True positives means the number of positive examples that the model correctly classified as positive, true negatives means the number of negative examples that the model correctly classified as negative, false positives means the number of negative examples that the model incorrectly classified as positive, and false negatives means the number of positive examples that the model incorrectly classified as negative. Sensitivity, specificity, and accuracy can be defined in Table 4. Both precision and recall must be examined to fully evaluate the effectiveness of any model. Precision and recall can be defined as in Table 4. F1-score is the measurement of the harmonic mean of recall and precision. This score considers both false positive and false negative values for calculation. For dataset-wise TP, TN, FP, and FN are computed using Table 4. We evaluated metrics, like sensitivity (recall), specificity, precision, and F1-score, and the classifiers, based on accuracy.

## 5 Results

The KNHANES is a national program that is designed to assess the health and nutritional status of adults and children in Korea. Obesity is defined as BMI above 25 according to the Asia-Pacific adult guidelines[8]. In our data, the obese people combined obese (above BMI 25) and overweight (above BMI 23 and below BMI 25). After removing missing values, the participants for males and females were around 21,100 (male 11,100, female 10,000). As can be seen from Figure 1, the ratio of obese people in the 2016-2018 years was 58% people, and in the 2019

**Table 4.** Some metrics for the evaluation of a classifier based on the confusion matrix.

| Metric | Equation |
|---|---|
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| Specificity | $\frac{TN}{FP+TN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Recall(sensitivity) | $\frac{TP}{TP+FN}$ |
| F1-score | $\frac{2\times Precision\times recall}{Precision+recall}$ |

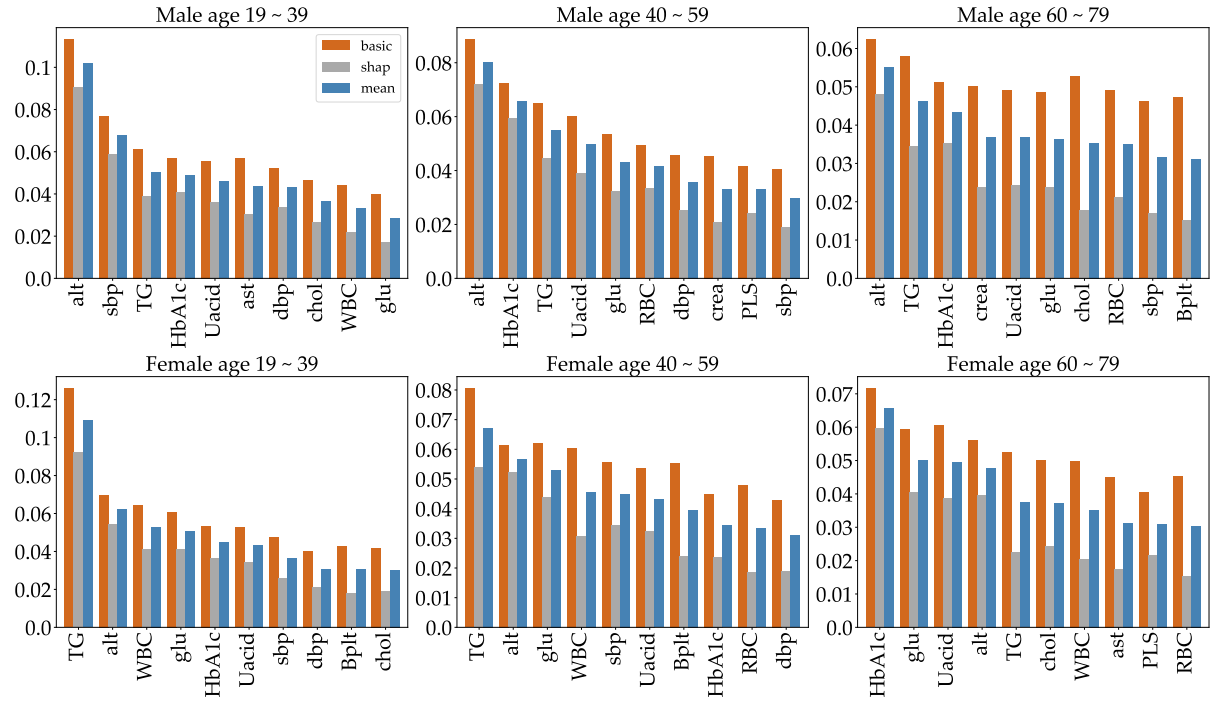


**Figure 4.** The top-10 features selected by age group and gender-specific. Orange bar indicates the feature importance using Random Forest, while gray bar shows the mean absolute Sharp value. Blue bar shows the mean of the two values.

year, the ratio was 57% people who were classified as obese above BMI 23. After we split the dataset into two parts by gender, we divided both the male and female into three age groups (19-39, 40-59, and 60-79), respectively. Therefore, our dataset consists of six groups composed according to gender and age group.

To predict obesity, it is very important to identify the risk factors of obesity. Also, we identified metabolic features that contribute to BMI or obesity under three different age-groups for each gender (hence, six datasets). Therefore, first, the metabolic variables were selected based on previous research to identify the principal features that contribute to BMI. Second, we carried out the correlation analysis between the metabolic factors in KNHANES and the target feature, BMI. Lastly, we used the mean of the Random Forest and mean absolute Sharp value, and selected ten factors for each of the six age-specific groups. The selected features made up a slightly different feature according to the six age groups. The 10 features were selected of 6 age groups as given in Figure 4. The features are ALT (SGPT), glucose (glu), Triglyceride (TG), Hemoglobin (Hb), White blood cell (WBC), Glycated hemoglobin (HbA1c), creatinine (cre), Systolic (diastolic) blood pressure (sbp/sdp), Ast (SGOT), Cholestol (chol), Hematocirit (Hct), Platelet (Bplt), and Uric acid (UA) (see Fig. 4).

For all three male groups, the ALT(SGPT) was the most important feature in both the Sharp value and Random Forest. For female, in two age groups, Triglyceride (TG) was the most important feature, except for the 60-79 age group. In the 60-79 age female group, HbA1c was the most important feature. The feature creatinine (crea) could only be selected from the 40-59 and 60-79 age male groups. Also, a special feature is the diastolic blood pressure, which was not selected for the 60-79 age groups of both male and female. Note that the SMOTE technique is used for oversampling due to class imbalance in the dataset. The number of obese and normal classes seems imbalanced. Table S1 of the supplementary information gives more detailed imbalance information of the six datasets. The SMOTE technique with oversampling was used, which resulted in two new datasets of 50% and 50%. Next, we employed eight different machine learning algorithms to predict age-specific and gender-specific obesity for adults (19-79 years) in Korea. The eight machine learning algorithms were performed using the top eight features as shown in Figure 4.

Furthermore, we measured the performance of these classifications in terms of several selected performance metrics. The Table S2 of SI illustrates commonly used performance metrics by machine learning methods. Figure 5 shows the ROC curves of the performance of algorithms for the three age groups per gender. The best performance classifier among the six age groups is the Multi-Layer Perceptron, which has the best performance for five age groups, except for the 60-79 age female group. Multi-Layer Perceptron achieved an area under the ROC curve (AUC) of 0.78 in the 19-39 age female group, while achieving an AUC of 0.77 in the 19-39 age male group. However, in the 60-79 age group and both male and female groups, MLP achieved a lower (AUC) of 0.7. The RF and LF achieved the same AUC of 0.66 in the 60-79 age female group (see Fig. 5). In particular, the MLP, RF, GBM, LGBM, and XGB achieved the same AUC of 0.72, while the LR achieved an AUC of 0.71 for the 40-59 age female group(see Table S4 of the SI). Those algorithms achieved better performance in each dataset. However, SVM and KNN performed poorly (see Table S3 and Figure S1 of the SI).

The accuracy of each classifier according to the 6 age groups differed for male and female, and the accuracy of female was higher than the accuracy of male for each age group, except the 40-59 age group. RF algorithm achieved the best accuracy in all six age groups, but for two groups of females and the male 60-79 age group, accomplished below 0.7 accuracy. In the case of male, RF and GBM achieved a higher 0.72 accuracy in the 40-59 age group, and RF achieved 0.72 accuracy in the 19-39 age group. In the 60-79 age group, of both male and female, all algorithms achieved lower 0.7 accuracy. Table S3 of the SI shows that KNN and SVM classifiers achieved lower accuracy for all age groups and gender. The Table S2 of SI lists the performance of each of the six algorithms. The performance ability is determined according to its accuracy, sensitivity, specificity, precision, recall, F1-score, and ROC-AUC for the six age groups. As a result of subjecting the six groups to the eight algorithms, the accuracy and ROC were the highest in the 19-39 age female group. The best algorithm for accuracy was RF, and for ROC the best were MLP and LR algorithms.

Lastly, as the features changed up to the top 10, the accuracy was measured by performing each algorithm for

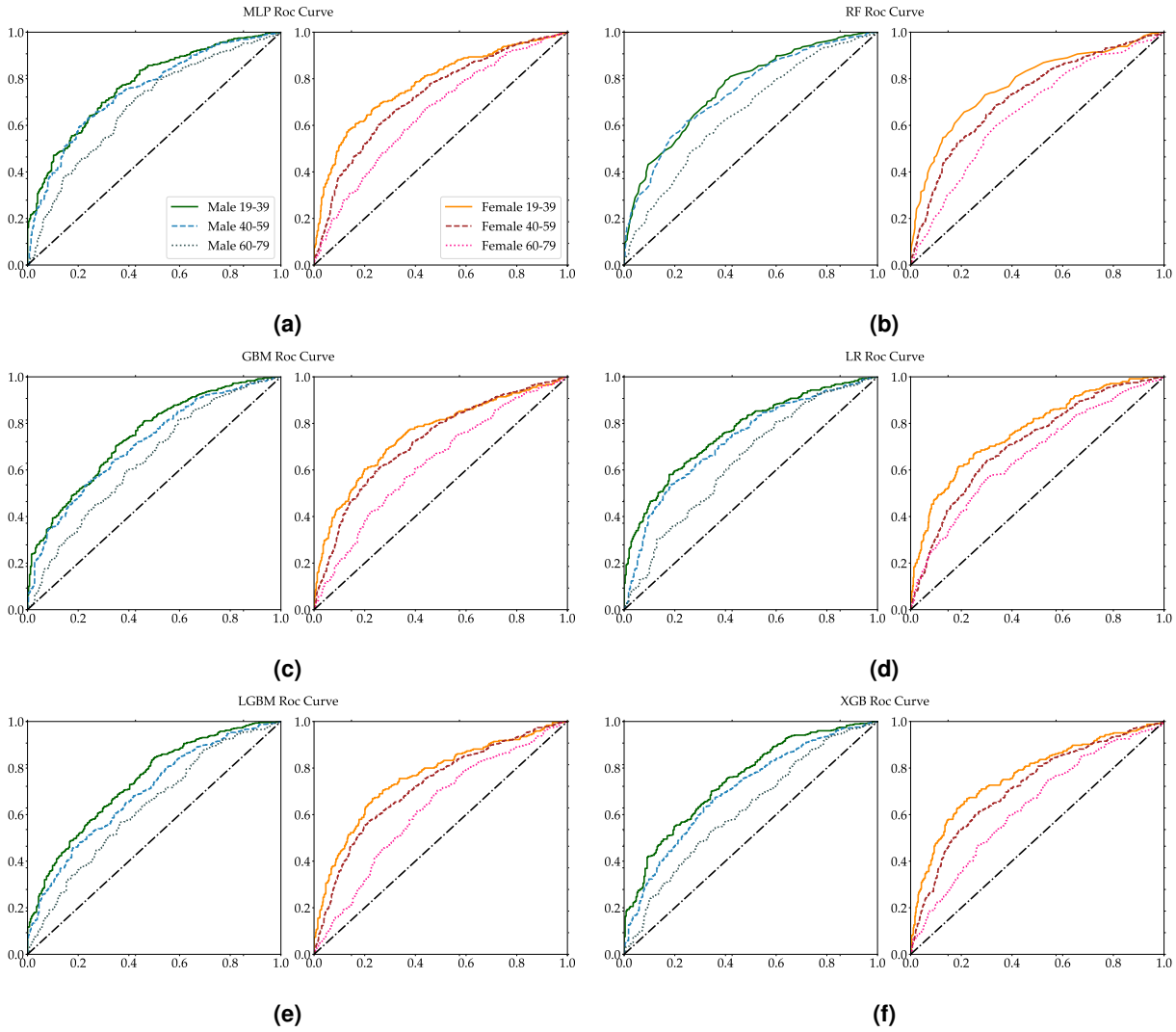


**Figure 5.** Age-specific ROC curves for each MLP, RF, GBM, LR, LGBM, and XGB. The left panel of each pair is male, while the right panel is female.

the six datasets (see Fig. 6). Even if the number of features is increased to (5, 6, 7, 8, 9, up to 10), we cannot obtain the relations with the accuracy and the number of features. Meanwhile, the accuracy of SVM tended to decrease in the female 19-39 age group. Also, the accuracy of algorithms shows little increase up to the top 5 or 6, but the KNN classifier decreases up to the top 5 features, except for the female 40-59 age group. In the case of males, the 19-39 and 40-59 age groups tend to show a slight increase in the RF, LGBM, and GBM algorithms. However, the algorithmic accuracy of the datasets did not change significantly depending on the number of the top 6 or more features.

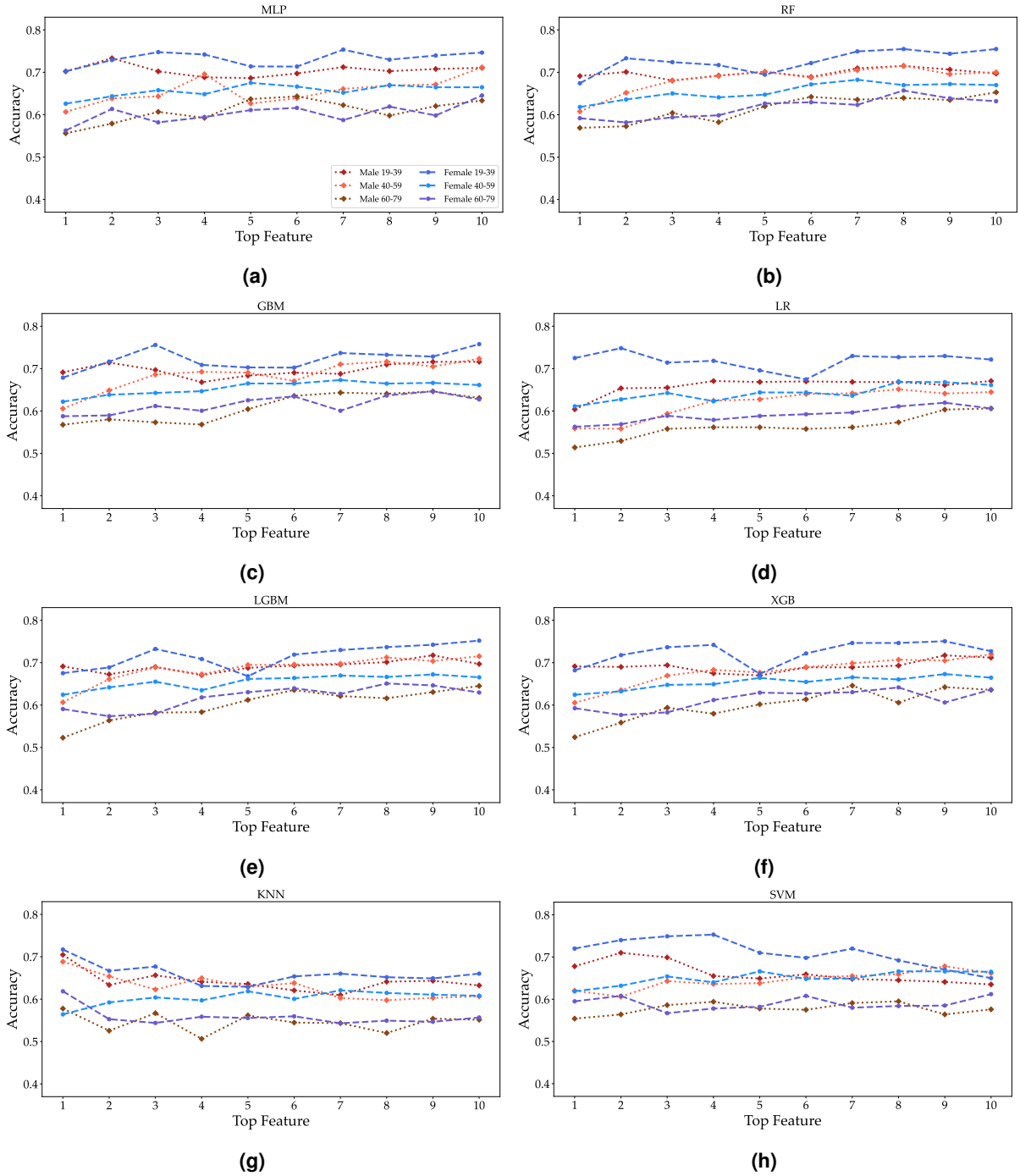


**Figure 6.** Accuracy is displayed as the number of features from 1 to 10 under three age groups and two genders. These features are extracted from the RF feature importance and mean absolute SHAP values in Fig. 4.

## 6 Discussion

This study provides identifying risk factors for obesity in adults among the extensive data set provided by the KDCA from 2016 to 2019 in Korea. First, in every 6 datasets, each of the top-10 features were selected, which are shown in Figure 4. The 10 features by the mean of random forest and mean absolute Shapley values include ALT(SGPT), glucose, Triglyceride, Hemoglobin, White blood cell, Glycated hemoglobin, creatinine, Systolic (diastolic) blood pressure, Ast(SGOT), Cholestol, Hematocrit, Platelet, and Uric acid. For the three male datasets, the ALT(SGPT) was the most important feature in the mean of Sharp value and Random Forest. For two female datasets, Triglyceride was the most important feature, except for the 60-79 age group. In the 60-79 age female group, HbA1c was the most important feature. The feature creatinine could only be selected for the male 40-59 and 60-79 age groups. In fact, there are other important factors that might influence obesity (e.g., food consumption, food production, physical activity, social psychology, genetic, physiological and cultural influences, etc.) that were not included for our analysis, since these were not suitable for a machine learning approach. Further research needs to be carried out using these factors with individual variances.

Well-known factors for predicting obesity are age, gender, waist circumference, and race. These factors state the consequences of obesity. Moreover, various factors that cause obesity have been studied, and there have been many predictive studies using factors that affect obesity[1,47,48]. Although a few researchers have included metabolic factors as key features in predicting obesity, most researchers have used individual factors, and individual lifestyle/behavioral and environmental factors as essential features[1,17,21]. We identified essential metabolic factors affecting obesity, and performed machine learning utilizing metabolic factors as key features. Our performance could not be compared with other research results, because the selected features are different[1,49,50].

Therefore, one of the main limitations of the study is that our features are not yet standard features for predicting obesity. It is important to evaluate and build predictive models for obesity using common risk factors. It would be desirable to improve standardized or common factors affecting obesity, because various works have dealt with different features. Another limitation of our model is the imbalanced data set from the KDCA, in which larger high-BMI population (people above BMI 23) than low-BMI population (people below BMI 23) was observed. To facilitate future research, common features to predict obesity are required, since there are many various factors that cause obesity. Hence, if standard factor assumptions are possible, common features to predict obesity may be used to carry out elaborate prediction and sophisticated mathematical analysis.

## Acknowledgements

C. Oh was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (2020R1I1A306562712). Sunmi Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2021R1A2B5B01002611).

## Author contributions statement

J. Jeon. conducted the experiment(s), S.Lee and C. Oh write this manuscript and analysed the results.

## Additional information

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://biomath.khu.ac.kr/index.php/resources/

## References


1. Safaei, M., Sundararajan, E. A., Driss, M., Boulila, W. & Shapi'i, A. A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Comput. biology medicine* **136**, 104754, DOI: https://doi.org/10.1016/j.compbiomed.2021.104754 (2021).

2. Butland, B. *et al. Tackling obesities: future choices-project report*, vol. 10 (Citeseer, 2007).

3. Collaborators, G. . O. Health effects of overweight and obesity in 195 countries over 25 years. *New Engl. J. Medicine* **377**, 13–27, DOI: https://doi.org/10.1056/NEJMoa1614362 (2017).

4. Jung, Y., Ko, S. & Lim, H. The socioeconomic cost of adolescent obesity. health and social welfare review 30: 195-219 (2010).

5. Yongik, K. Socioeconomic effects of obesity. *Natl. Heal. Insur. Serv. korea* (2018).

6. Müller-Riemenschneider, F., Reinhold, T., Berghöfer, A. & Willich, S. N. Health-economic burden of obesity in europe. *Eur. journal epidemiology* **23**, 499–509, DOI: https://doi.org/10.1007/s10654-008-9239-1 (2008).

7. von Lengerke, T. & Krauth, C. Economic costs of adult obesity: a review of recent european studies with a focus on subgroup-specific costs. *Maturitas* **69**, 220–229, DOI: https://doi.org/10.1016/j.maturitas.2011.04.005 (2011).

8. Lim, J. U. *et al.* Comparison of world health organization and asia-pacific body mass index classifications in copd patients. *Int. journal chronic obstructive pulmonary disease* **12**, 2465, DOI: https://doi.org/10.2147/COPD.S141295 (2017).

9. Organization, W. H. *et al.* The asia-pacific perspective: redefining obesity and its treatment. (2000).

10. https://knhanes.kdca.go.kr/knhanes.

11. Lakshmanaprabu, S. *et al.* Online clinical decision support system using optimal deep neural networks. *Appl. Soft Comput.* **81**, 105487, DOI: https://doi.org/https://doi.org/10.1016/j.asoc.2019.105487 (2019).

12. Vijayarani, S. & Dhayanand, S. Liver disease prediction using svm and naïve bayes algorithms. *Int. J. Sci. Eng. Technol. Res. (IJSETR)* **4**, 816–820 (2015).

13. Kim, E. *et al.* Application of machine learning to predict weight loss in overweight, and obese patients on korean medicine weight management program. *The J. Korean Medicine* **41**, 58–79, DOI: https://doi.org/10.13048/jkm.20015 (2020).

14. López-Martínez, F., Núñez-Valdez, E. R., Crespo, R. G. & García-Díaz, V. An artificial neural network approach for predicting hypertension using nhanes data. *Sci. Reports* **10**, 1–14, DOI: https://doi.org/10.1038/s41598-020-67640-z (2020).

15. Cheng, X. *et al.* Does physical activity predict obesity—a machine learning and statistical method-based analysis. *Int. J. Environ. Res. Public Heal.* **18**, 3966, DOI: https://doi.org/10.3390/ijerph18083966 (2021).

16. Thamrin, S. A., Arsyad, D. S., Kuswanto, H., Lawi, A. & Nasir, S. Predicting obesity in adults using machine learning techniques: An analysis of indonesian basic health research 2018. *Front. nutrition* **8**, DOI: https://doi.org/10.3389/fnut.2021.669155publisher={FrontiersMediaSA} (2021).

17. Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I. & Habib, M. T. A machine learning approach for obesity risk prediction. *Curr. Res. Behav. Sci.* **2**, 100053, DOI: https://doi.org/10.1016/j.crbeha.2021.100053 (2021).

18. Chatterjee, A., Gerdes, M. W. & Martinez, S. G. Identification of risk factors associated with obesity and overweight—a machine learning overview. *Sensors* **20**, 2734, DOI: https://doi.org/10.3390/s20092734 (2020).

**19.** Singh, B. & Tawfik, H. Machine learning approach for the early prediction of the risk of overweight and obesity in young people. In *International Conference on Computational Science*, 523–535, DOI: https://doi.org/10.1007/978-3-030-50423-6_39 (2020).

**20.** Dunstan, J. *et al.* Predicting nationwide obesity from food sales using machine learning. *Heal. informatics journal* **26**, 652–663, DOI: https://doi.org/10.1177/1460458219845959 (2020).

**21.** Jindal, K., Baliyan, N. & Rana, P. S. Obesity prediction using ensemble machine learning approaches. In *Recent Findings in Intelligent Computing Techniques*, 355–362, DOI: https://doi.org/10.1007/978-981-10-8636-6_37 (2018).

**22.** Montañez, C. A. C. *et al.* Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 2743–2750, DOI: https://doi.org/10.1109/IJCNN.2017.7966194 (2017).

**23.** Grabner, M. Bmi trends, socioeconomic status, and the choice of dataset. *Obes. facts* **5**, 112–126, DOI: https://doi.org/10.1159/000337018 (2012).

**24.** Kim, Y. The korea national health and nutrition examination survey (knhanes): current status and challenges. *Epidemiol. health* **36**, DOI: https://doi.org/10.4178/epih/e2014002 (2014).

**25.** Kweon, S. *et al.* Data resource profile: the korea national health and nutrition examination survey (knhanes). *Int. journal epidemiology* **43**, 69–77, DOI: https://doi.org/10.1093/ije/dyt228 (2014).

**26.** Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *J. artificial intelligence research* **16**, 321–357 (2002).

**27.** Ko, K. & Oh, C. The development of an obesity index model as a complement to bmi for adult: Using the blood data of knhanes. *Honam Math. J.* **43**, 717–739 (2021).

**28.** Yen Jean, M.-C. *et al.* Association between lifestyle and hematological parameters: A study of chinese male steelworkers. *J. Clin. Lab. Analysis* **33**, e22946, DOI: https://doi.org/10.1002/jcla.22946 (2019).

**29.** Landi, F. *et al.* Body mass index is strongly associated with hypertension: Results from the longevity check-up 7+ study. *Nutrients* **10**, 1976, DOI: https://doi.org/10.3390/nu10121976 (2018).

**30.** Akter, R., Nessa, A., Sarker, D. & Yesmin, M. Effect of obesity on hemoglobin concentration. *Mymensingh Med. Journal: MMJ* **26**, 230–234 (2017).

**31.** Adams, L. A., Knuiman, M. W., Divitini, M. L. & Olynyk, J. K. Body mass index is a stronger predictor of alanine aminotransaminase levels than alcohol consumption. *J. gastroenterology hepatology* **23**, 1089–1093, DOI: https://doi.org/10.1111/j.1440-1746.2008.05451.x (2008).

**32.** Stranges, S. *et al.* Body fat distribution, relative weight, and liver enzyme levels: A population-based study. *Hepatology* **39**, 754–763, DOI: https://doi.org/10.1002/hep.20149 (2004).

**33.** Joshi, S., Godbole, G. *et al.* Correlation of body mass index & triglyceride levels in middle aged women. *Atherosclerosis* **275**, e227, DOI: https://doi.org/10.1016/j. (2018).

**34.** Zou, Y., Sheng, G., Yu, M. & Xie, G. The association between triglycerides and ectopic fat obesity: an inverted u-shaped curve. *PloS one* **15**, e0243068, DOI: https://doi.org/10.1371/journal.pone.0243068 (2020).

**35.** Akter, R. *et al.* Effect of obesity on fasting blood sugar. *Mymensingh medical journal: MMJ* **26**, 7–11 (2017).

**36.** Pratley, R. E., Wilson, C. & Bogardus, C. Relation of the white blood cell count to obesity and insulin resistance: effect of race and gender. *Obes. research* **3**, 563–571, DOI: https://doi.org/10.1002/j.1550-8528.1995.tb00191.x (1995).

**37.** Das, R., Nessa, A., Hossain, M., Siddiqui, N. & Hussain, M. Fasting serum glucose and glycosylated hemoglobin level in obesity. *Mymensingh Med. Journal: MMJ* **23**, 221–228 (2014).

**38.** Banfi, G. & Del Fabbro, M. Relation between serum creatinine and body mass index in elite athletes of different sport disciplines. *Br. journal sports medicine* **40**, 675–678, DOI: https://doi.org/10.1136/bjsm.2006.026658 (2006).

**39.** Chang, J.-B. *et al.* The role of uric acid for predicting future metabolic syndrome and type 2 diabetes in older people. *The journal nutrition, health & aging* **21**, 329–335, DOI: https://doi.org/10.1007/s12603-016-0749-3 (2017).

**40.** Chu, F.-Y. *et al.* The association of uric acid calculi with obesity, prediabetes, type 2 diabetes mellitus, and hypertension. *BioMed Res. Int.* **2017**, DOI: https://doi.org/10.1155/2017/7523960 (2017).

**41.** Samocha-Bonet, D. *et al.* Platelet counts and platelet activation markers in obese subjects. *Mediat. inflammation* **2008**, DOI: https://doi.org/10.1155/2008/834153 (2008).

**42.** Purdy, J. C. & Shatzel, J. J. The hematologic consequences of obesity. *Eur. J. Haematol.* **106**, 306–319, DOI: https://doi.org/10.1111/ejh.13560 (2021).

**43.** Marchesini, G., Moscatiello, S., Di Domizio, S. & Forlani, G. Obesity-associated liver disease. *The J. Clin. Endocrinol. & Metab.* **93**, s74–s80, DOI: https://doi.org/10.1002/hep.23280 (2008).

**44.** Faheem, M. *et al.* Does bmi affect cholesterol, sugar, and blood pressure in general population? *J. Ayub Med. Coll. Abbottabad* **22**, 74–77 (2010).

**45.** Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. neural information processing systems* **30** (2017).

**46.** Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. neural information processing systems* **30** (2017).

**47.** DeGregory, K. *et al.* A review of machine learning in obesity. *Obes. reviews* **19**, 668–685, DOI: https://doi.org/10.1111/obr.12667 (2018).

**48.** Colmenarejo, G. Machine learning models to predict childhood and adolescent obesity: a review. *Nutrients* **12**, 2466, DOI: https://doi.org/10.3390/nu12082466 (2020).

**49.** Golino, H. F. *et al.* Predicting increased blood pressure using machine learning. *J. obesity* **2014**, DOI: https://doi.org/10.1155/2014/637635 (2014).

**50.** Zheng, Z. & Ruggiero, K. Using machine learning to predict obesity in high school students. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2132–2138, DOI: https://doi.org/10.1109/BIBM.2017.8217988 (2017).

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.pdf