# Comparing Performance of Ensemble-Based Machine Learning Algorithms to Identify Potential Obesity Risk Factors from Public Health Datasets


**Ayan Chatterjee, Martin W. Gerdes, Andreas Prinz, and Santiago G. Martinez**

**Abstract** Societal factors such as globalization, supermarket growth, rapid unplanned urbanization, sedentary lifestyle, economical distribution, and social position gradually develop behavioral risk factors in humans. Behavioral risk factors are unhealthy habits (consumption of tobacco and alcohol), improper diet (consumption of high calorific discretionary fast foods, sweet beverages), and physical inactivity. The behavioral risks may lead to physiological risks, body–energy imbalance. Obesity is one of the foremost lifestyle diseases that leads to other health conditions, such as cardiovascular disease (CVDs), chronic obstructive pulmonary disease (COPD), cancer, diabetes type II, hypertension, and depression. It is not restricted within the boundary of age and socio-economic background. "World health organization (WHO)" has predicted that lifestyle diseases will claim 71–73% of the global death, by the end of 2020. It can be prevented with proper identification of associated risk factors and appropriate behavioral intervention plans. The key determinants of obesity are—a. age, b. weight, c. height, and d. body mass index (BMI). This paper addresses the potential of ensemble machine learning approaches to assess the associated risk factors of obesity through the evaluation of existing, publicly accessible health datasets, such as "Kaggle", and "UCI". Followed by, we compared our identified risk factors with the obtained risk factors from literature study. In future, we are intending to reuse the obtained knowledge to collect data from a controlled trial

A. Chatterjee (✉) · M. W. Gerdes · A. Prinz
Department of Information and Communication Technology, Centre for e-Health, University of Agder, 4604 Kristiansand, Norway
e-mail: ayan.chatterjee@uia.no

M. W. Gerdes
e-mail: martin.gerdes@uia.no

A. Prinz
e-mail: andreas.prinz@uia.no

S. G. Martinez
Department of Health and Nursing Science, Centre for e-Health, University of Agder, 4604 Kristiansand, Norway
e-mail: santiago.martinez@uia.no



A. E. Hassanien et al. (eds.), *Emerging Technologies in Data Mining and Information Security*, Advances in Intelligent Systems and Computing 1286,
https://doi.org/10.1007/978-981-15-9927-9_26

of adult population (age between 20 and 60) in south Norway to generate personalized, contextual, and behavioral recommendations with a smart electronic coaching (eCoaching) system for behavioral intervention for the promotion of healthy lifestyle.

**Keywords** Ensemble-based algorithm · Machine learning algorithms · Electronic coaching

## 1 Introduction

In 1980, 1.3 million people were obese globally, and the number doubled in 2008. Unhealthy habits (such as tobacco consumption, alcoholic beverages), unhealthy diet (such as energy drinks, consumption of excess salt and sugar, intake of high saturated fat, and discretionary foods), physical inactivity are the major pillars of obesity. In 2016, more than 1.9 billion adults (39%), aged 18 years and older, were overweight, and of these, over 650 million (13%) were obese. In 2016, more than 340 million children and teenagers aged 15–19 were obese, and in 2018, 40 million children under the age of 5 were obese. The universal predominance of obesity nearly tripled between 1975 and 2016. Juvenile obesity is linked to a higher chance of obesity, premature death, and infirmity in adulthood [1–4].

Obesity that continues as the foremost public health anxiety increases the risk of other four primary lifestyle diseases such as cardiovascular diseases (CVD), cancers, diabetes (type II), and chronic lung diseases (COPD, asthma). The burden of these diseases is extremely high among 63% (36 million) of global death occurred in 2008 due to lifestyle diseases. 80% of the 36 million dead people belonged to low- and middle-income classes, 13% were from high-income classes, and 29% of total lifestyle-related deaths occurred below the age of 60 years. In 2016, the number increased to 56.9 million (71%), and by 2030, it is predicted to achieve 75% with 88.5% death in the developed countries and 65% death in the developing countries. The risk for lifestyle disease increases with body mass index (BMI) [1–5]. Health behavior change should be given the highest precedence to circumvent severe damages.

Electronic coaching (eCoaching) system can empower people to manage a healthy lifestyle with early risk predictions and appropriate individualized recommendations. To develop an intelligent eCoach system for automated, personalized, contextual, behavioral recommendations to achieve personal wellness goals, addressing obesity as a study case, we need to a. identify associated health risk factors and b. collect them for further analysis with ensemble learning methods [1, 2].

The main contributions of this paper are as follows—(a) Identification of a set of risk factors associated with obesity following statistical approaches on datasets available in “Kaggle”, and “UCI” and (b) Analysis of different ensemble learning methods for classification on the same selected datasets.

The remainder of the paper is structured as follows. In Sect. 2, we summarized the related works. Section 3 presents the methodology utilized. In Sect. 4, we discussed

our findings. The paper is concluded in Sect. 5. We excluded concepts such as, "child obesity", "pregnancy", "genetics", and "eCoach" [6, 7] from this study. We focused only on the adult population with age >20 and age <60. This study is an extended research work of the paper entitled with "Identification of Risk Factors Associated with Obesity and Overweight-A Machine Learning Overview" [8].

## 2 Related Work

Obesity remains a significant public health problem not only in the USA but also in other countries for the last 10–15 years. It has prevailed among pre-school students and childbearing-age females at a low rate but is increasing among school students rapidly and remains high in adults, mainly in the group of females with less education or schooling. In developed countries, it occurs mostly in the group of weak and the opposite occurs in less developed societies as household nutrition transition and underweight can coexist with the weight increase. Obesity tends to decline with increased income. In developed countries, females are suffering almost double of males in the lower socioeconomic group [3].

Different projects have been conducted by different research groups on "obesity-related risk predictions" with machine learning approaches to generate prediction and classification models. Singh et al. [2] evaluated different multivariate regression methods and multilayer perceptron (MLP) feed-forward neural network models on the dataset obtained from a millennium cohort study (MCS) with over 90% accuracy to predict the teenager BMI from previous BMI values. Twenty neurons in the hidden layer resulted in the lowest mean absolute error (MAE) with a mean training time of 1.63 s and a regularization factor of 0.9. Bassam et al. [9] performed a study on data obtained from the Kuwait Health Net-work (KHN) to build prognostic models to predict the future risk of diabetes (type II) using machine learning algorithms (logistic regression, K-nearest neighbor, support vector machine) with five-fold cross-validation technique. The study included age, sex, body mass index (BMI), pre-existing hypertension, family history of hypertension, and diabetes (type II) as baseline non-invasive parameters. As a result, K-NN outperformed the other models with AUC values of 0.83, 0.82, and 0.79 for 3-, 5-, and 7-year prediction limits. Meghana et al. [9] used "auto-sklearn", an automatic machine learning (AutoML) library for developing classifiers of CVDs. They experimented on both the heart-UCI dataset and cardiovascular disease dataset consist of 70,000 records of patients, and as a result, AutoML outperformed traditional machine learning classifiers. Seyla et al. [10] studied how to classify obesity from dietary and physical activity patterns using machine learning classification algorithms! and as a result, support vector machine (SVM) outperforms other classifiers.

Jindal et al. [11] performed ensemble-based machine learning approaches for obesity prediction based on the key determinants, such as—age, height, weight, and BMI. The ensemble model utilized random forest (RF), generalized linear model, and partial least square with a prediction accuracy of 89.68%. Grabner

et al. performed a study on "National Health and Nutrition Examination Survey (NHANES)", "National Health Interview Survey (NHIS)", and "Behavioral Risk Factor Surveillance System (BRFSS)" datasets from the 1970s to 2008, to analyze the trend of BMI in the USA over time, across race/ethnicity, gender, and socioeconomic status (SES) groups, and across different datasets. It was observed that SES-BMI gradients were steadily more significant for females than males.

## 3 Methodology

The focus of our research is to perform statistical analysis on available datasets in "Kaggle" and "UCI" to identify a list of potential risk factors associated with obesity. Subsequently, evaluate the performance of different ensemble methods for multi-class classification analysis. The overall process includes traditional well-established machine learning methodologies—a. data collection, b. data pre-processing, c. statistical analysis and data visualization, d. algorithm selection for classification and feature predictions, e. model training and testing, and f. model evaluation [12, 13].

### *3.1 Data Collection*

The datasets for this study are selected from "Kaggle" [14] and "UCI" [15]. Selected datasets are summarized in Table 1. We have targeted three types of datasets—a. obesity, b. CVDs and c. diabetes as CVDs and diabetes (type II) are closely related to the risk of obesity.

**Table 1** Selected datasets for statistical analysis and ensemble learning [8]

| Repository | Name | Sample size | Source | Category |
|---|---|---|---|---|
| Kaggle | 500_Person_Gender_Height_Weight_Index (BMI) | 500 | www.github.com | Obesity |
| Kaggle | Insurance | 1338 | www.cseastbay.edu | Obesity |
| Kaggle | Eating-health-module-dataset [16] | 11,212 | US Bureau of Labor Statistics | Obesity |
| Kaggle/UCI | Pima-Indians-diabetes-database | 768 | UCI Machine Learning | Diabetes |
| Kaggle | Cardiovascular-disease-dataset | 462 | Ryerson University | CVDs |

**Table 2** Python libraries for data processing [14]

| No. | Libraries | Purpose |
|---|---|---|
| 1 | Pandas | Data importing, structuring, and analysis |
| 2 | NumPy | Computing with multidimensional array object |
| 3 | Matplotlib | Python 2D plotting |
| 4 | SciPy | Statistical analysis |
| 5 | Seaborn, plotly | Plotting of high-level statistical graphs |
| 6 | Scikit-learn (sklearn) | Machine learning, preprocessing, cross-validation, and evaluate model's performance |
| 7 | Graph Viz | Plotting of decision trees |

## *3.2 Data Processing*

Accumulated data in this research are labeled. Therefore, we have used ensemble learning classification models for training and testing the accuracy. Some selected datasets are small; some are noisy, and the remaining are containing a good volume of data to train an ensemble machine learning model. Data mining included to filter data samples from each of the datasets and to discard samples containing outliers. Data mining involves pattern discovery, calculation of feature correlation, feature selection, and classification.

Data processing incorporates three steps as stated below: [12–14]

a. Data preprocessing includes—a. data integration, b. removal of noisy data that are incomplete and inconsistent, c. data normalization and feature scaling, d. encoding of categorical data, e. feature selection after correlation analysis, and f. split data for training and testing an ensemble machine learning model.
b. Training of learning model and test its accuracy with k-fold cross-validation.
c. Data postprocessing includes a. pattern evaluation, b. pattern selection, c. pattern interpretation, and d. pattern visualization.

In this experiment, we have used python 3.x programming language libraries described in Table 2. We set up a "Python" environment using "Anaconda Distribution" and used "Spyder IDE".

## *3.3 Statistical Analysis*

Statistical analysis of the selected datasets involves the following methods, as stated in Table 3. According to the central limit theorem, when a bunch of random numbers is added together, it produces a normal distribution. The normal distribution can be described entirely by the two parameters $\mu$ (mean) and $\sigma$ (standard deviation). As always, the mean is the center of the distribution, and the standard deviation is the measure of the variation around the mean. Probability of normal distribution can be

**Table 3** Hypothesis testing methods [13, 14]

| Method | Description | Samples |
|---|---|---|
| T test | Test if mean of a normally distributed value is different from a specified value ($\mu_0$) | Sample size < 30 |
| Z test | Test if two samples are equal or not | Sample size > 30 |
| ANOVA or F test | Test multiple groups at the same time | More than 2 samples |
| Chi-Square test | Check if observed patterns (O) of data fit some given distribution (E) or not | Two categorical variables from a sample |

calculated through the standard normal distribution "Z":

$$|\boldsymbol{Z}| = \left| \frac{\boldsymbol{X} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \right|$$

The Z-score transformation is a linear transformation with $\mu = 0$ and $\sigma = 1$. It is used for feature scaling. A normality test is used to check if a distribution is Gaussian or not. The normal distribution is symmetric about $\mu$. So, the area to the left of $\mu$ is equal to the area to the right of $\mu$.

We have used hypothesis testing as a statistical analysis in this study. A hypothesis test estimates two mutually exclusive statements about a population to ascertain which statement is best supported by the trial data. The critical parameter of hypothesis testing is the null hypothesis ($\mathbf{H_0}$), and the alternative hypothesis (**Ha**) that directly contradicts $\mathbf{H_0}$. The confidence factor ($\alpha$) is used to decide whether to accept or reject an $\mathbf{H_0}$. The value of "$\alpha$" is usually kept as 0.05 or 5%, as 100% accuracy is impossible to achieve whether to accept or reject an $\mathbf{H_0}$. Popular, widely used hypothesis testing methods, a short description, and the required sample size are demonstrated in Table 3. A hypothesis test can be either a one-tailed test or a two-tailed test. For each of the testing methods, resulting probability value (P-value) is compared with "$\alpha$" to accept or reject a null hypothesis. But it suffers from type-I error (false positive) or type-II error (false negative) [13, 14].

"Shapiro–Wilk", "D'Agostino's K^2", "Anderson–Darling" test calculate P-value to decide if a sample looks like Gaussian (P-value > $\alpha = 0.05$) or not (P-value < $\alpha = 0.05$). Covariance (COV(x, y)) is a property of a function to retain its form when its variables are linearly transformed. It helps to measure correlation ($r_{xy}$) that measures the strength of the linear relationship between two variables.

$$corr(x, y) = COV(x, y)/(\sigma x * \sigma y), \quad where \ -1 < r < +1$$

The "Sign" of "r" shows the direction of the relationship among two variables x and y. Table 4 shows the meaning of different |r| values. If two variables/features are strongly correlated, then we selected any one of them during feature selection. Pearson's correlation coefficient is used to summarize the strength of the linear

**Table 4** Statistical analysis methods on the selected datasets [17]

| \|r\| value | Meaning |
|---|---|
| 0.00–0.2 | Very weak |
| 0.2–0.4 | Weak to moderate |
| 0.4–0.6 | Medium to substantial |
| 0.6–0.8 | Very strong |
| 0.8–1 | Extremely strong |

**Table 5** Statistical analysis methods on the selected datasets [8, 13, 14]

| No | Methods | Purpose |
|---|---|---|
| 1 | Mean, standard deviation, skewness | Distribution test |
| 2 | t-test, z-test, F-test, Chi-square | Hypothesis test |
| 3 | Shapiro–Wilk, D'Agostino's K^2, and Anderson–Darling test | Normality test |
| 4 | Covariance, correlation | Association test |
| 5 | Histogram, Swarm, Violin, Bee Swarm, Joint, Box, Scatter | Distribution plot |

relationship between two variables in normal distribution and spearman's correlation is used to calculate the non-linear relationship between two variables [12–14] (Table 5).

## *3.4 Model Training and Testing*

In this study, we have selected the following ensemble learning methods, such as bagging, boosting, and voting for multi-class classification analysis as described in Table 6.

The description of the ensemble learning methods is described as follows [11, 14]—a. Bagging/Bootstrap Aggregation: It works well with the algorithms that have high variance, such as decision tree (DT), random forest (RF). In this method, the ensemble model tries to improve prediction accuracy and decrease model variance by combining predictions of individual models trained over randomly generated

**Table 6** Ensemble learning methods [11–14]

| Ensemble learning methods | Ensemble algorithms |
|---|---|
| Bagging | Bagged decision tree |
| Bagging | Random forest |
| Bagging | Extra trees |
| Boosting | AdaBoost |
| Boosting | Stochastic gradient boosting |
| Voting | Voting ensemble |

training samples. The final prediction of the ensemble model is given by calculating the average of all predictions from the individual estimators, b. Boosting: It combines several weak base learners, trained sequentially over multiple iterations of training data, to build powerful ensemble. During the training of weak base learners, higher weights are assigned to those learners which were misclassified earlier, and c. Voting: In this method, multiple models of different types are constructed on some straightforward statistical methods, such as mean or median and they are used to combine the predictions. This prediction serves as the additional input for training to make the final prediction.

In our study, for the "voting" ensemble-based classification method we used seven classifiers as follows—(a.) SVM with kernel "RBF" or non-linear kernel, (b.) SVM with linear kernel, (c.) Gaussian NB, (d.) Decision Tree Classifier, (e.) Random Forest Classifier, (f.) KNeighbors Classifier, and (g.) Linear Discriminant Analysis. For other methods, such as "Bagging", and "Boosting" we used decision tree models with n_estimators = [50, 100, 150, 200], and random_state = 7. The best value of the "n_estimators" was obtained following the "grid-search" method.

The steps used to train and test an ensemble machine learning model are described below: [8]

- Load data
- Data pre-processing following the below steps:
  - remove missing value from the loaded data
  - encode categorical features
  - check distribution of data and features
  - correlation analysis among features and feature scaling if required
  - shuffle the data
- Split data for training and testing (80:20) with some random state
- Ensemble learning method selection as described in Table 6 based on classification problem statement
- K-fold cross-validation on data (in our study, K = 5)
- Evaluate model performance with metrics as described in Sect. 3.5
- Perform model tuning with "grid search" parameter optimization technique where required.

**N.B:** a. The selection of learning rate ($\alpha$): if too small then slow convergence in gradient descent (GD) and if too large then slow convergence in GD or GD may diverge.
b. Let; "m" training samples have "n" features. If, too many features ($m <= n$), then delete some features or use regularization with regularization factor "$\lambda$".
c. If "$\lambda$" is too large then the algorithm fails to eliminate overfitting, or even sometimes underfit and GD fails to converge. "$\lambda$" ($\infty$) increases to lead high bias and decreases to lead high variance.
d. Underfitting results in high bias and overfitting leads to high variance.

g. Gradient descent follows convex optimization technique with upper bound (L) and lower bound (μ) on curvature f:

$$\mu I_d \leq \nabla^2 f(x) \leq L I_d, \ where \ \nabla^2 f(x) \ is \ the \ Hessian, \ \mu > 0 \ and \ L = Lipschitzcontinious$$

### 3.5 Model Evaluation

Developed ensemble learning methods for classification are evaluated with below metrics [8, 12–14].

Classification metrics: accuracy score, classification report, and confusion matrix. Classification report includes precision, recall and F1-score. A confusion matrix is a table with two dimensions "actual" and "predicted" and both the dimensions have "true positives (TP)", "true negatives (TN)", "false positives (FP)", "false negatives (FN)".

- TP—both actual class and predicted class of data point is 1.
- TN—both actual class and predicted class of data point is 0.
- FP—actual class of data point is 0 and predicted class of data point is 1.
- FN—actual class of data point is 1 and predicted class of data point is 0.

Formulas for calculating classification metricss are stated as below:

$$Accuracy = \frac{(TP + TN)}{TP + FP + FN + TN}, Precision(P) = \frac{TP}{(TP + FN)},$$
$$Recall \ (R) \ or \ Sensitivity(S) = \frac{TP}{TP + FN},$$
$$Specificity = (1 - Sensitivity) = \frac{TN}{TN + FP}, F1score = 2 * \frac{PR}{P + R}$$

Accuracy tells how close a measured value is to the actual one. Precision determines how close a measured value is to the actual one. Recall or sensitivity defines the total number of positives (actual) returned by the ensemble machine learning model.

### 3.6 Assessment of Body Composition

BMI has been used to categorize different weight groups in adults of 20 years or older, both male and female: [4, 8]

- Underweight: BMI < 18.5
- Normal weight: BMI is 18.5–24.9

- Overweight: BMI is 25–29.9
- Obese: BMI is 30 or more.

## 4 Results and Discussion

Analysis of "BMI" dataset with 500 records reveals that index (body composition) has a strong correlation with BMI as depicted in Fig. 1. BMI column was externally added during data pre-processing and later, was removed during model training due to high correlation. The dataset has six classes for classification under index field—extremely weak, weak, normal, overweight, obesity, and extreme obesity.

In multiclass classification, the "voting"-based ensemble outperformed other methods as described in Table 6 with an accuracy score of 89% as depicted in Fig. 2.

We added extra feature "body_composition" to the "insurance" dataset with 1338 records based on the BMI feature, and the feature classifies the records among four classes—underweight, normal weight, overweight, and obese. We encoded the categorical features such as sex, smoker, and region. There is a strong correlation between "smoking" and "charge" with $|r| = 0.79$. Smoking grows negative health behavior in humans. Negative health behavior has a great impact on weight change, obesity, and overweight. So, excess smoking does not only create an active negative impact on health but also creates a passive negative impact on economic position. We have used insurance data for multi-class classification analysis.

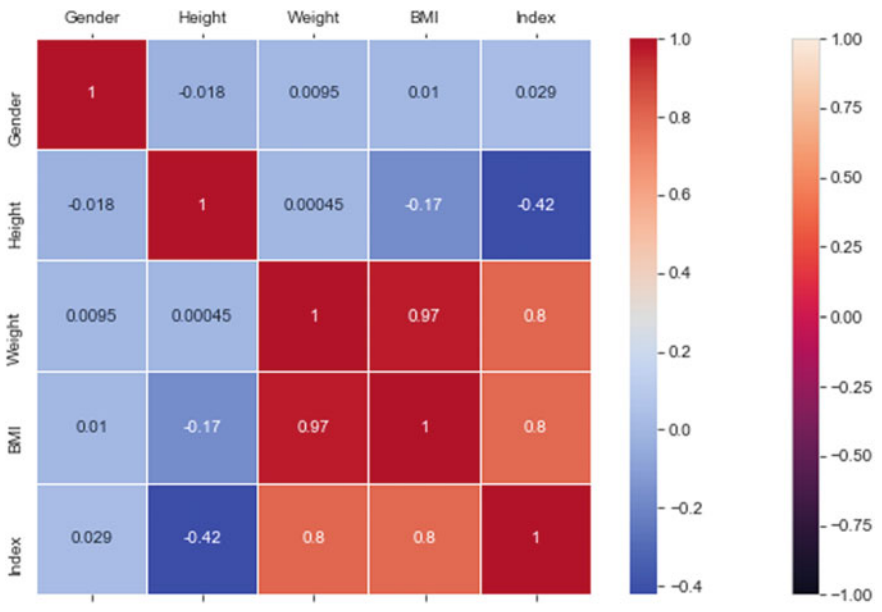


**Fig. 1** Correlation heatmap of "BMI" data [8]

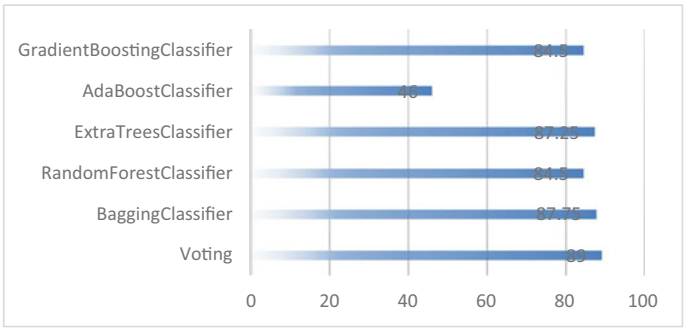


**Fig. 2** Classification accuracy (%) of ensemble learning models to classify 'BMI' data

During classification, we used "body_composition" as a predicted feature, and the accuracy of ensemble learning models is depicted in Fig. 2 and "GradientBoostingClassifier" has performed the best with 99.9% accuracy, n_estimators = 100, random_state = 7, following grid search method (Fig. 3).

For the regression, we used "charges" as predicted feature and performed hypothesis testing with "ANOVA" results to retain Ha = {a significant change between the three age categories (young adults, senior adults, elders) with "BMI"} with a P-value of 0.001, 0.060, and 0.000 respectively.

The boxplot (left) in Fig. 4 demonstrates that the obesity risk increases with age, and the mean BMI for the three-age categories is in the obesity range, which is a risk. The regression plot (right) in the same figure depicts how insurance charge increases with obesity. Insurance charge also increases with smoking condition and age as depicted in Fig. 5, and Fig. 6, respectively.

"Eating-health-module-dataset" with 11,212 records was processed with ensemble ML classification algorithms to classify records in between four classes "underweight", "normal weight", "overweight", and "obese" under "body_composition" feature and the "BaggingClassifier", "AdaBoostClassifier", and

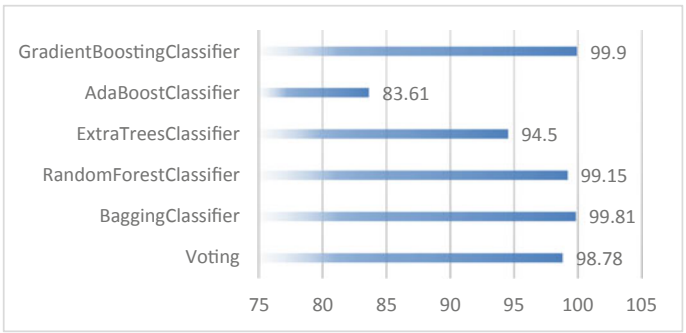


**Fig. 3** Classification accuracy of ensemble ML models to classify "body_composition" insurance data

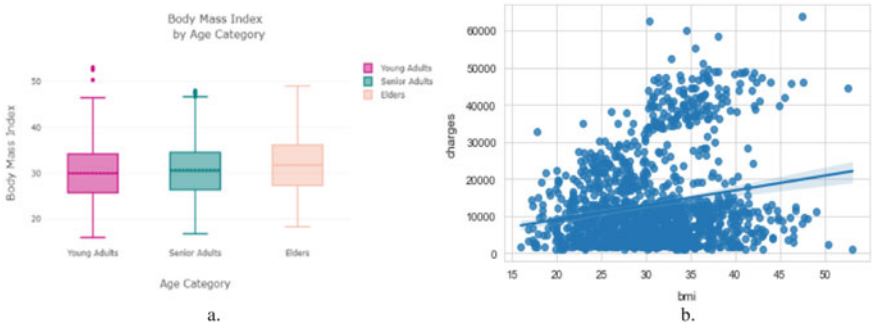


**Fig. 4** **a** Relationship in between "Age category" and "BMI" [8]; **b** Relationship in between "BMI" and "charges" [8]

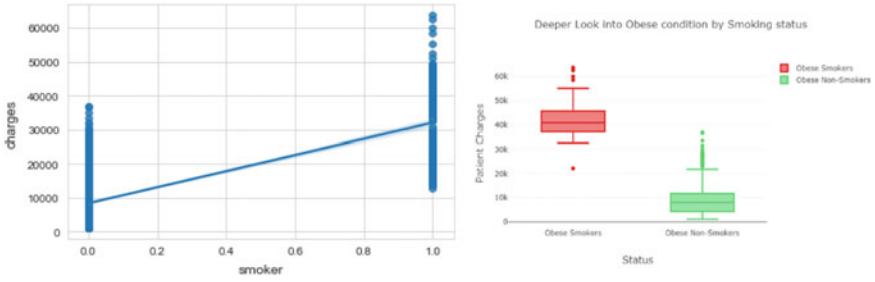


**Fig. 5** Relationship in between "charges" and smoking condition

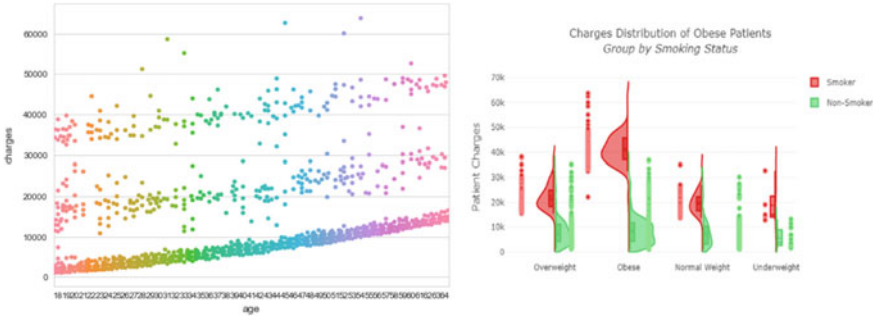


**Fig. 6** Relationship in between "charges" and "age"

"GradientBoostingClassifier" classifiers performed the best as depicted in Fig. 7 with accuracy = 99.9%, n_estimators = 50, and random_state = 7, following grid search method.

The regression analysis in Fig. 8 shows that sweet beverages, economic condition, fast food, sleeping, meat and milk consumption, drinking habit, exercise has a sharp

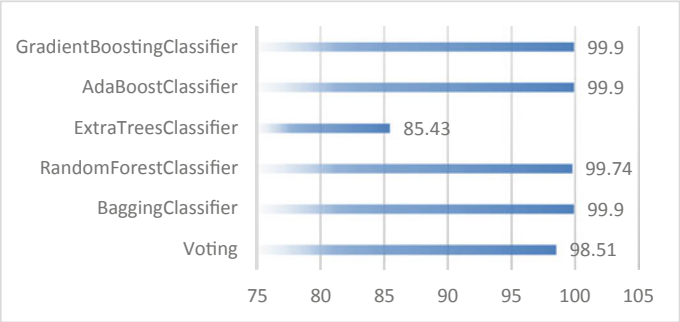


**Fig. 7** Classification accuracy of ensemble ML models to classify "body_composition" data

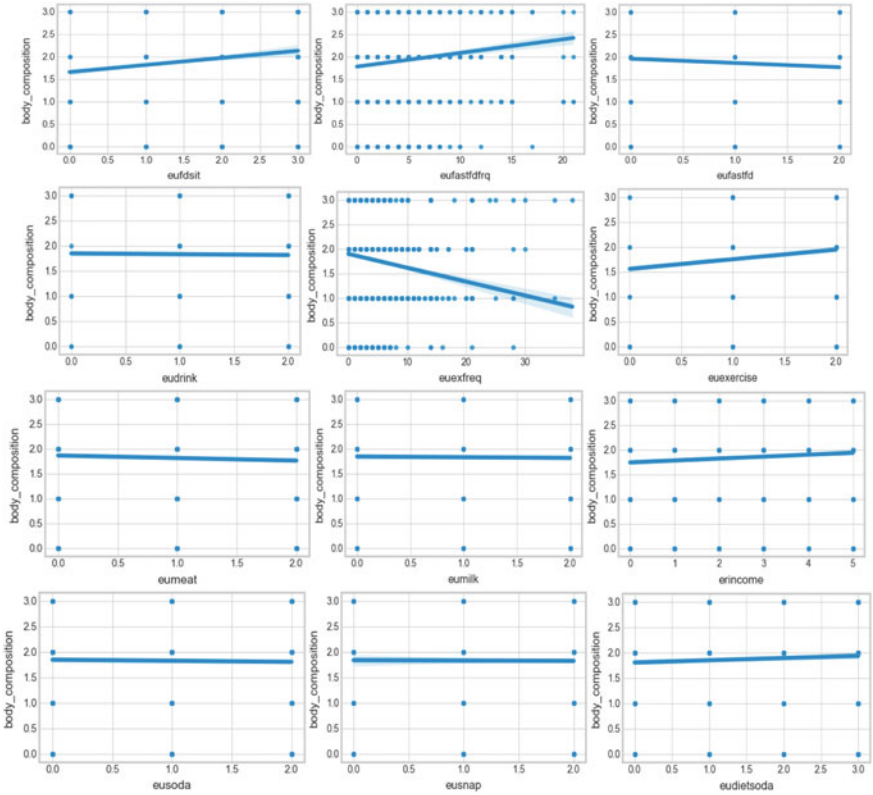


**Fig. 8** Regression analysis of selected parameters with "body_composition" in eating health module dataset [18]

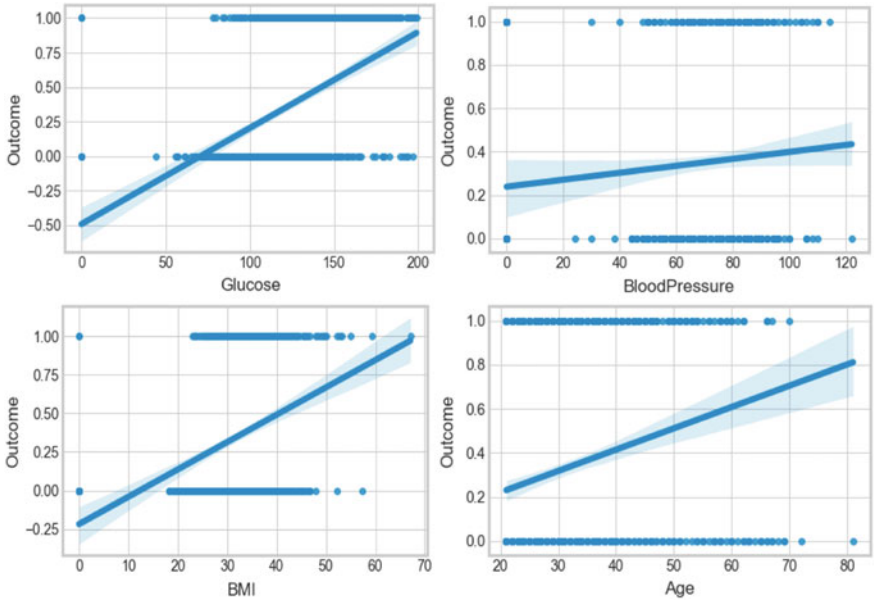


**Fig. 9** Regression analysis of diabetes dataset to show a positive relationship between weight change and age, BMI, blood glucose, and blood pressure

impact on growing obesity in human. The regression analysis of "Pima-Indians-diabetes-database" dataset with 768 records resulted in a positive dependency on between weight change and discovered factors such as blood glucose, blood pressure, and age as depicted in Fig. 9.

We used ensemble ML classification algorithms to classify records among two classes "obese" (1) and "non-obese" (0) under feature column "outcome" and "Voting"-based ensemble method outperformed other classifiers as depicted in Fig. 10. This analysis shows a strong relationship between obesity and diabetes.

The regression analysis of "cardiovascular-disease-dataset" with 462 records shows, in Fig. 11, that blood pressure, tobacco consumption, lipid profile, adiposity, family history, obesity, drinking habit, and age have a strong connection with CVDs. In binary classification problems on the used heart dataset, "GradientBoostingClassifier" outperformed other classifiers with accuracy = 70.7%, n_estimators = 50, and random_state = 7, following grid search method as depicted in Fig. 12.

From the above data analyses, we observed how different ensemble-based machine learning classification methods are performing on different publicly available health datasets! The "Voting", and "Gradient Boosting" ensemble classification machine learning method offered consistent performance on every dataset. The identified risk factors of obesity from the above analyses can be summarized as—a. BMI, b. age, c. tobacco consumption, d. sweet beverages, e. economic condition, f. fast food, g. sleeping pattern, h. diet, i. blood pressure, j. blood glucose, k. lipid profile, l. adiposity, m. exercise, and n. family history.

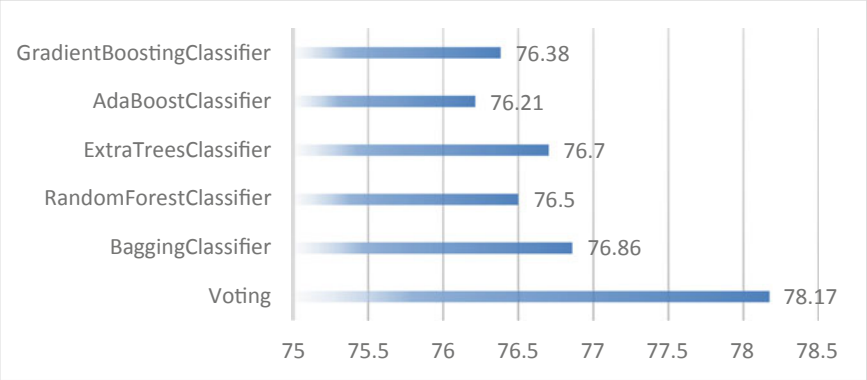


**Fig. 10** Classification accuracy of ensemble ML models to classify diabetes dataset

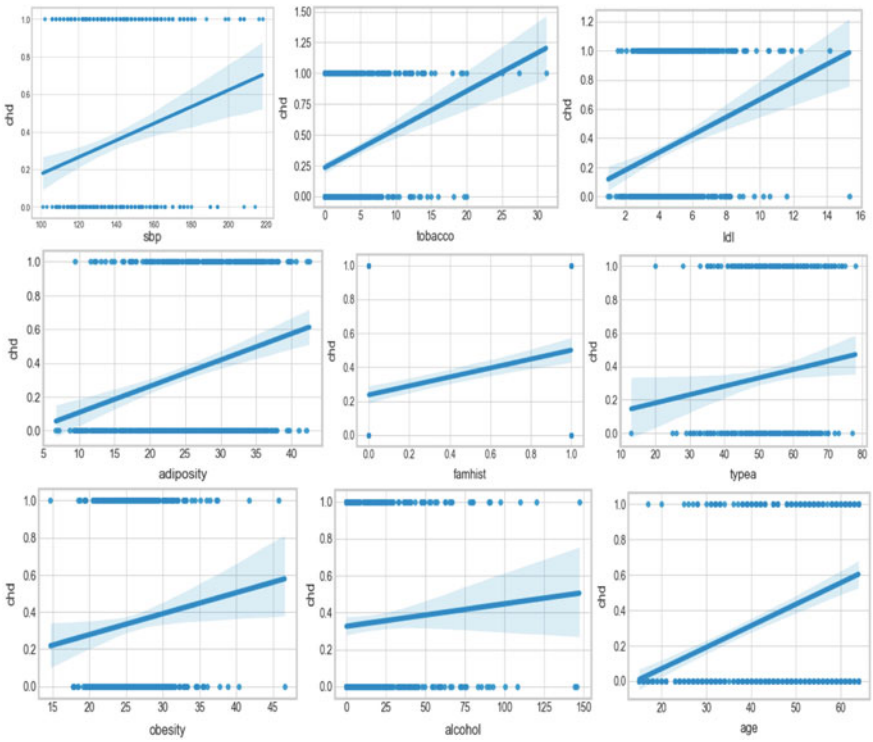


**Fig. 11** Regression analysis of cardiovascular disease dataset

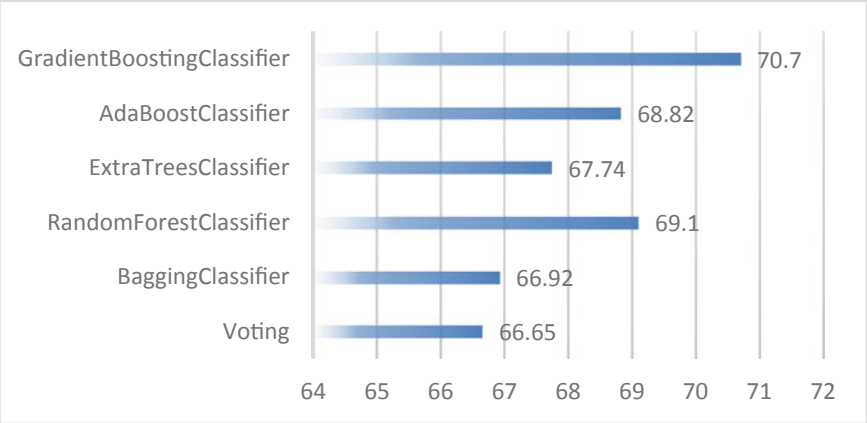


**Fig. 12** Classification accuracy of ensemble ML models to classify cardiovascular dataset

## 5 Conclusion

This study aims to identify probable risk factors associated with obesity, after studying existing health datasets, publicly available in "Kaggle", and "UCI" with statistical and ensemble learning methods. The trouble we found with the size of data, target population, and data integration. The statistical analysis has given a clear indication of the potential risk factors to be addressed and further studied. In the future, the study can be extended with non-convex optimization (artificial neural net, deep learning). Our future study aims to design, develop, test, and evaluate the performance of an intelligent eCoach system for automatic generation of personalized, contextual behavioral recommendations to address health and wellness challenges related to obesity. For the same, we will collect health data related to associated risk factors from controlled participants over time, as identified from this study.

## References

1. Butler, É.M., et al.: Prediction models for early childhood obesity: applicability and existing issues. In: Hormone Research in Paediatrics, pp. 358–367 (2018)
2. Singh, B., Tawfik, H.: A machine learning approach for predicting weight gain risks in young adults. In: 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT), pp. 231–234 IEEE (2019)
3. Grabner, M.: BMI trends, socioeconomic status, and the choice of dataset. In: Obesity Facts, pp. 112–126 (2012)
4. WHO page. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
5. Csige, I., Ujvárosy, D., Szabó, Z., Lőrincz, I., Paragh, G., Harangi, M., Somodi, S.: The impact of obesity on the cardiovascular system. J. Diabet. Res. (2018)

6. Gerdes, M., Martinez, S., Tjondronegoro, D.: Conceptualization of a personalized ecoach for wellness promotion. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 365–374 (2017)
7. Chatterjee, A., Gerdes, M.W., Martinez, S.: eHealth initiatives for the promotion of healthy lifestyle and allied implementation difficulties. In: 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1–8. IEEE (2019)
8. Chatterjee, A., Gerdes, M.W., Martinez, S.G.: Identification of risk factors associated with obesity and overweight—a machine learning overview. Sensors **20**(9), 2734 (2020)
9. Padmanabhan, M., Yuan, P., Chada, G., Van Nguyen, H.: Physician-friendly machine learning: a case study with cardiovascular disease risk prediction. J. Clin. Med., 1050 (2019)
10. Selya, A.S., Anshutz, D.: Machine learning for the classification of obesity from dietary and physical activity patterns. In: Advanced Data Analytics in Health, pp. 77–97. Springer, Cham (2018)
11. Jindal, K., Baliyan, N., Rana, P.S.: Obesity prediction using ensemble machine learning approaches. In: Recent Findings in Intelligent Computing Techniques, pp. 355–362. Singapore (2018)
12. Schapire, R.E., Freund, Y.: Boosting: foundations and algorithms. In: Kybernetes (2013)
13. Brandt, S.: Statistical and computational methods in data analysis. No. 04; QA273, B73 1976. In: Amsterdam: North-Holland Publishing Company (1976)
14. Sklearn page. https://scikit-learn.org/stable/supervised_learning.html
15. Kaggle data page. https://www.kaggle.com/data
16. Eating-health-module-dataset description. https://www.bls.gov/tus/ehmintcodebk1416.pdf
17. Chatterjee, A., Gerdes, M.W., Martinez, S.G.: Statistical explorations and univariatetimeseries Analysis on COVID-19 datasets to understand the trend of disease spreading and death. Sensors **20**(11), 3089 (2020)
18. Python page. https://docs.python.org/