Determining Practical Estimators of Body Fat Percentage

Hunter McCue
hmccue@wisc.edu

Rachel Rothenberg
rrothenberg@wisc.edu

Casey Lin
cmlin@wisc.edu

Abstract

Body mass index, or BMI, is a measurement of the body used to define if a person is overweight, underweight, or in a healthy weight category. Calculating the BMI relies on the mass and the height of the individual.

$$BMI = mass/height^2$$

While there are many ways to estimate the body mass of a person, calculating an accurate measure can be a difficult and expensive task. We proposed using machine learning models in order to come up with an accurate calculation for the BMI of a male person using measurements of their body, hoping to find a more convenient way for doctors to get access to accurate BMI calculations. We use 7 models to predict BMI, and compare them based on their mean squared error and R squared values, using Linear Regression as the baseline. Among the 7 models tested, we found that Gradient Boosting and Random Forests were able to deliver performance on par with Linear Regression, with Extra Random Trees beating out Linear Regression in both metrics. We find that these tree-based ensemble methods as viable alternatives to linear regression for BMI prediction.

1. Introduction

1.1. Background

Our goal in this project is to compare multiple different machine learning techniques in order to find which ones are able to accurately calculate the BMI of an individual. In our data set, we are given several measurements of the body such as age, neck circumference, hip circumference, and more. Compared to the current methods to accurately calculate BMI, simple body measurements are significantly simpler to obtain. These measurements do not require the use of expensive equipment or underwater testings, making them more convenient to obtain. Inputting these factors into our machine learning models, we aim to predict the BMI of the person with high accuracy.

Rather than only trying and tuning one model, we use several and compare their accuracy to one another. This allows us to truly find the best fitting model for our given data set, and rule out the ones that have higher rates of error.

1.2. Motivation

Body fat is able to help medical professionals learn more about a person's health, and sometimes predict what problems they may have in the future. A BMI that is too low can lead a person to be more susceptible to illnesses, and potentially have lower energy levels as well. A BMI that is too high can lead to health issues such as diabetes, heart conditions, and higher rates of having severe reactions to infections and viruses. Similar studies, which will be discussed in detail in the next section, have discovered that having a BMI which is deemed unhealthy can lead to further risk of severity and fatality of COVID-19 if developed. Additionally, a large BMI can be the underlying cause of being at risk of many other diseases including heart disease, hypertension, type 2 diabetes, gallstones and some types of cancer. This further demonstrates the need for a practical estimator of BMI, as the measurement is extremely useful and can potentially determine if individuals are at higher risk of serious diseases or ailments.

Accurate calculations of BMI are often expensive and inconvenient to find. Some of the most accurate ways to calculate BMI include hard-to-find body scanners, underwater weighing, or expensive air displacement machines. In most cases, these methods of measuring BMI are inaccessible, which leads doctors to either use inaccurate predictions of their BMI, or to omit it entirely. Finding a way to accurately calculate BMI based on just body measurements will allow doctors to have access to more in depth information about the patient and to better prepare for the potential health problems that individual may have.

1.3. Applications

If we are able to create a model that can accurately predict a person's BMI based on simple measurements of their body, we can allow doctors to gain accurate insight into a patient's health without the need for inconvenient and expensive testing. This could allow patients to be more aware of health issues they may face in the future, and potentially prepare and take steps in reducing their risk. Knowing a

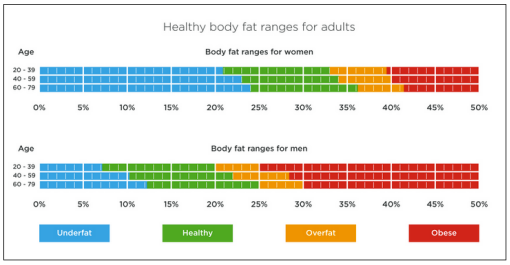

Figure 1. Body fat percentage can be used as an indicator of one's healthiness, however, it is difficult to measure practically. [2]

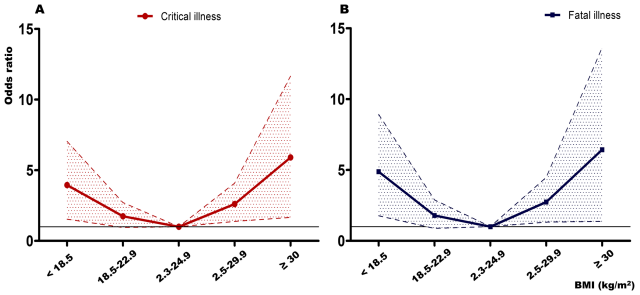

Figure 2. Results from the similar experiment demonstrates the difference in odds for critical/fatal illness for different BMI intervals [5]

person's BMI allows the doctor to tell if the patient is at a healthy weight for their height, and can help them decide if lifestyle changes are necessary in order for the patient to remain healthy. Knowing ahead of time what potential health risks may arise can allow the patient to make the needed changes in order to reduce their risk of developing these health issues before they become too big of a problem.

2. Related Work

A previous study was performed pertaining 2,656 middle-aged individuals (40-69 years old) whose body fat percentage and BMI were recorded [4]. Individuals were then classified into different health groups based on each factor and a chi-square test was performed to see if there was a difference between the two groupings. The grouping of the individuals by body fat percentage was inconsistent with what resulted from the grouping of BMI, with 30 percent of the body fat percentage classification not being amiable with the BMI results. Therefore, it was concluded that body fat percentage was a poor predictor as it could falsely classify someone as having a healthy weight when they actually are overweight or obese based on their BMI. This further shows the importance of having a practical and simple way to measure one's BMI, as it eliminates the need to use slightly inaccurate classifiers such as body fat percentage.

A study was conducted which analyzed 18,940 individuals who had contracted COVID-19 sometime in the first four months of 2020 [6]. The individuals were divided into 4 groups depending on their BMI level: underweight, normal weight, overweight and obese. Each group was analyzed to determine if there was a higher risk of COVID-19 for a specific group using logistic regression analysis. The findings showed that there was an increased risk of developing COVID-19 with higher BMIs, with overweight individuals being about 1.13 times more likely to develop COVID-19 and obese individuals being about 1.26 times more likely to develop COVID-19 than normal weight individuals.

Another similar study was performed to look at the relationship between BMI and COVID-19. 4,141 patients of COVID-19 were examined with data provided from the Korean Centers for Disease Control and Prevention Agency. The severity and fatality of each patient's COVID-19 symptoms were measured and then compared to the individuals' BMI. Multiple logistic regression was used to statistically analyze the data. The results of the study found that there was a non-linear, U-shaped relationship between BMI and the fatality of COVID-19. Patients that had a BMI of less than 18.5 kilograms per square meter or a BMI greater than 25 kilograms per square meter were determined to have a higher fatality rate from COVID-19 than those that maintained a healthy BMI.

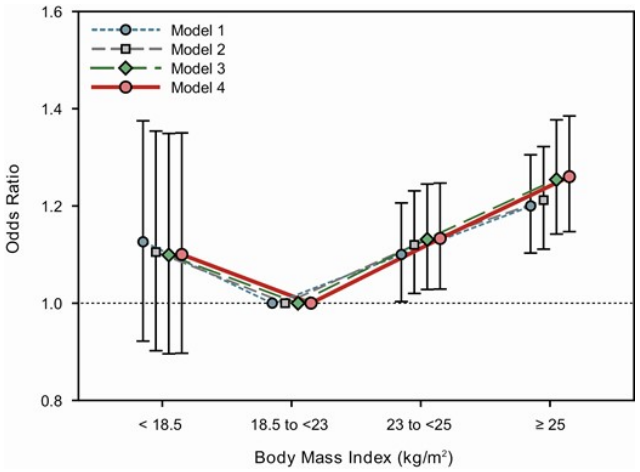

Figure 3. Throughout all 4 of the differently adjusted models, those outside of the healthy BMI interval are consistently more at odds of infection from COVID-19. [6]

3. Proposed Method

3.1. Problem Description and Methodology

We want to predict the percentage of body fat given the features and measurements in the data set, which is a regression problem. In order to compare machine learning methods to conventional methods, we use a linear regression model as our baseline. After that, we then fit multiple

machine learning methods(e.g. K-Nearest Neighbors, Decision Trees, Boosted Trees) to the data, which allows us to compare said methods with the conventional linear model as well as each other.

3.2. Linear Regression

Linear regression fits a straight line that minimizes the loss the most through the data. We use Linear Regression as our baseline model to compare the other models to, as it is the simplest regression model, takes very little time to solve, and has a closed-form solution.

3.3. K-Nearest Neighbors

K-Nearest Neighbors takes a data-point, and looks at the k nearest neighbors in the feature space, determined by a distance metric d. For regression, the algorithm then averages the values of said neighbors as the output.

3.4. Decision Tree

Decision trees uses nodes in order to classify each data point. Decision trees are simple to interpret, and don't require a lot of data preparation and cleaning. At each node step in the tree, it aims to separate the data into more specific and more accurate classifications. Additionally, a maximum depth is determined in order to prevent over fitting of the model.

3.5. Boosting and Ensemble Methods

Models with just one tree may not give a good fit, so often ensemble methods and boosting methods are used to help improve the performance of the models. This is done so by attempting to prevent outlier predictions as well as overfitting of the model. Additionally, these methods can be used to boost weak learners into becoming strong learners.

3.5.1 Random Forests

The Random Forest algorithm utilizes bagging with decision trees but the trees are split on random feature subsets rather than the complete feature set. Trees are fit onto different bootstrap samples and at each node of the tree random subsets of features are chosen to obtain the optimal split.

3.5.2 Extra Trees

The Extremely Randomized Trees algorithm, better known as the ExtraTrees algorithm is similar to random forests but at each decision tree node, a random feature is chosen for splitting. Therefore, this method works very fast due to the lack of computing information gain for each step in the tree.

3.5.3 AdaBoost

Adaptive boosting is one method that looks to boost weak learners into strong learners. It looks to do this by adjusting the weights of misclassified as well as correctly classified training examples, prioritizing misclassified examples more for later iterations.

3.5.4 Gradient Boosting

Gradient boosting, similar to adaptive boosting, also looks to boost weak learners into strong learners. In the gradient boosting process, base trees are created and further trees are then constructed based on errors from the earlier trees. This method is fairly expensive to train but produces strong results.

4. Experiments

4.1. Dataset

The dataset we used was from Kaggle []. 14 selected measurements were taken from a sample of 252 men, as well as the true body fat percentage.

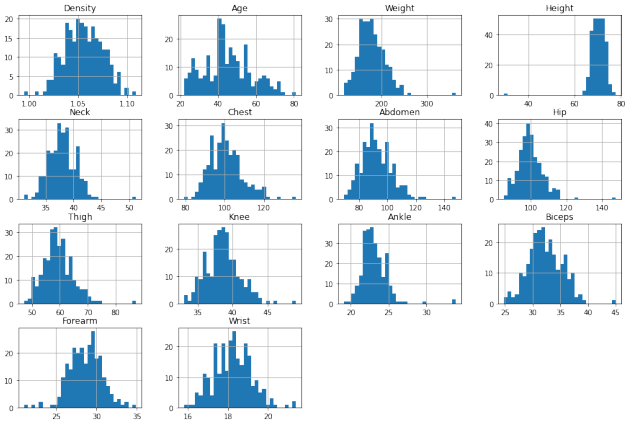

Figure 4. Distributions of the 14 features

4.2. Preprocessing Steps

4.2.1 K-Nearest Neighbors

For K-Nearest Neighbors, the efficacy of the model often depends on the scale of the data citation needed, so we used sci-kit learn's MinMaxScaler to normalize the data to $[0, 1]$ before training the model.

4.2.2 Other Methods

Because most of the proposed methods(Linear Regression, Tree-based methods) do not decrease in efficacy when the data is scaled, we didn't need to apply preprocessing to the data as the dataset was already cleaned for us.

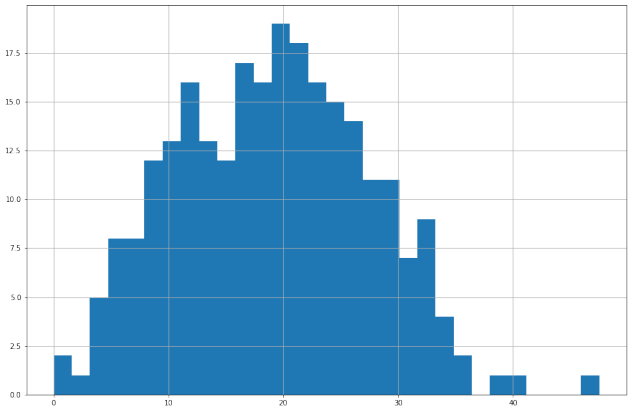

Figure 5. Distribution of body fat percentage

4.3. Software

Data Source: Our data set on Body Fat Percentage was obtained from Kaggle.

Data Preparation and Model Creation: We used Google Colab in order to write and run code on our data set.

Documentation: Overleaf and Google Documents were used for result presentation and documentation purposes.

Environments: We trained our models on the default hardware provided by Google Colab which is a Python developmental environment.

4.4. Hardware

Each member of the group used their own laptop which was either utilizing a Windows or Mac operating system.

5. Metrics Discussion

5.1. Metrics

We used 2 metrics, R^2 and Mean-Squared Error to assess the accuracy of the model. R^2 can also be defined as the coefficient of determination. It is a useful statistic that measures the proportion of variance of a dependent variable that's attributed to an independent variable. The Mean-Squared Error, also known as the MSE of an estimator is used to measure the squared difference of errors between predicted values from the model and actual values from the data. Therefore, the MSE can be used to determine the conditional quality of an estimator.

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

Method	R^2	MSE(95% CI)
Linear Regression	0.9720	1.9149 ± 38.8811
Decision Tree	0.9592	2.905 ± 33.8659
Gradient Boosting	0.9642	2.3365 ± 34.0617
Random Forests	0.9694	2.0773 ± 32.4747
Extra Trees	0.9793	1.3889 ± 19.6746
AdaBoost	0.9478	3.3868 ± 30.8587
K-Nearest Neighbors	0.8138	8.584 ± 29.3045

Table 1. Methods and Accuracies

Method	MSE(95% CI)(SE $>$ 50 removed)
Linear Regression	0.4027 ± 3.3369
Decision Tree	0.3922 ± 4.7484
Gradient Boosting	0.6321 ± 7.3683
Random Forests	0.4189 ± 5.3786
Extra Trees	0.3173 ± 4.2095
AdaBoost	1.4665 ± 7.8618
K-Nearest Neighbors	7.2627 ± 18.7242

Table 2. Accuracies with Outliers Removed

5.2. Validation

We used Leave-One-Out Cross Validation to assess the accuracy of the model, since the dataset was relatively small, with only about 200 datapoints. Since R^2 as a metric is undefined for only 1 sample, we had to first predict over all points and then find the R^2 over all predictions, which results in only 1 value for R^2. [3]

5.3. Training

When training the models, we used GridSearch with 20-fold CV for tuning the hyperparameters of the models. Because of the problems later mentioned with assessing the accuracy of R^2 with Leave-On-Out CV, we were unable to use Leave-One-Out CV to find the hyperparameters.

6. Results and Discussion

6.1. Training

All models were trained using GridSearchCV with 20 folds, since the score method of all models used used R^2, which was impossible to use with Leave-One-Out Cross Validation.

6.2. Confidence Intervals

In the results table, we have extremely large confidence intervals. This is due to certain testing examples that the models make large errors on, which inflates the MSE scores and standard deviation. However, examples that cause these outsized MSE values vary between model to model, so it would be impossible to find examples that consistently

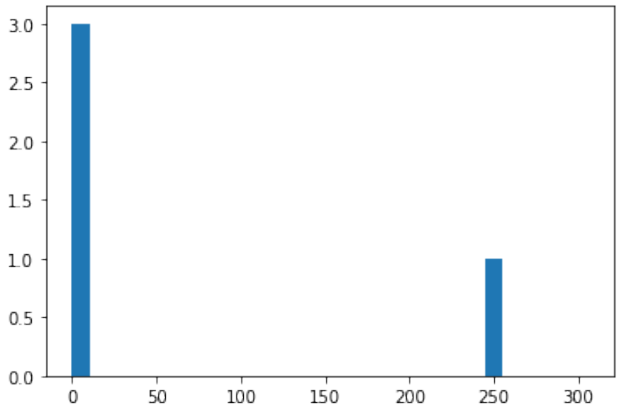

Figure 6. Distribution of Squared Errors for Linear Regression

cause these outsized MSE values. Because of this, we have included another metric, a 95% confidence interval for MSE with MSE values over 50 taken out. This metric should not be used to assess the true accuracy of the models but rather to assess the performance of the models when given non-outlier examples. More data may help with lessening the effect of these outliers on the MSE or may help the model perform better on these outlier examples.

6.3. Models

6.3.1 Linear Regression

The linear regression model was able to perform very well on our data set, achieving an R^2 value of .9720. This value was higher than the R^2 value for all of our other models. This high level of linear correlation is likely because there is a positive correlation between higher body measurements and higher BMI. This allowed the linear regression model to be able to fit well on our data, leading it to also have one of the lower Mean Squared Errors both with and without outliers.

6.3.2 K-Nearest Neighbors

For K-Nearest Neighbors, we applied Min-Max Scaling to the input features before prediction. After grid-search, we found that the best k was 5, with the euclidean distance metric. Furthermore, the model also performed better with distance weighting, where the model weights points by the inverse of their distance, allowing for closer points to have more influence [7] K-Nearest Neighbors had by far the worst performance on our dataset, having a MSE of 8.584 ± 29.3045 during validation. This may be a consequence of the small number of training examples in our dataset, as smaller training examples often leads to higher variance when making predictions. The model may be improved by getting a dataset with more training examples.

6.3.3 Decision Tree

GridSearch found that the decision tree performed best without max-depth pruning. The resulting decision tree had a MSE of 2.905 ± 33.8659, which performed worse than the boosting and ensemble methods, which was about expected. Because there was no pruning on the best estimator it is likely that the model may have been overfit, leading to worse performance during cross-validation.

6.3.4 Gradient Boosting

For Gradient Boosting, we found that relatively small trees with depth 2 and 50 estimators performed the best on the dataset. The performance of Gradient Boosting in general was comparable to Linear Regression, with a R^2 of 0.9642 and a MSE of 2.3365 ± 34.0617.

6.4. Random Forests

For Random Forests, no depth pruning and 100 estimators performed the best. The performance of the Random Forests was also comparable to Linear Regression and Gradient Boosting, with a R^2 of 0.9694 and a MSE of 2.077 ± 32.4747.

6.4.1 Extra Random Trees

Extra Random Trees had by far the best performance on our dataset, with the lowest average MSE and a much smaller confidence interval, 1.3889 ± 19.6746, compared to the other models, as well as having the highest R^2 at 0.9793.

6.4.2 AdaBoost

AdaBoost had the best hyperparameters at a learning rate of 0.25, with 25 estimators. However, AdaBoost did relatively poor compared to the other ensemble methods and even the singuler decision tree model, with a R^2 of 0.9478 and a MSE of 3.3868 ± 30.8587.

7. Conclusions

In this project, we aimed to fit machine learning models to our data set that would accurately predict the BMI of a man given a series of body measurements as input features. We fit several models to our data, including a linear regression model as our baseline to compare other models to. Based on the results, Extra Random Trees had the best performance on our model, being the only model to beat our baseline Linear Regression model. However, other ensemble and boosting methods such as Gradient Boosting and Random Forests also offered performance similar to the linear regression model. Thus, ensemble and boosting models

involving decision trees can be looked to as a viable alternative to the simple linear regression model. Furthermore, alternative models like K-Nearest Neighbors do well enough on the dataset to merit possibly further investigation with a larger dataset in the future. The biggest improvement that could be made in the future would be to get an alternative dataset with more training examples to help offset the effect of outlier training examples during validation.

We believe that with further testing and model tuning, along with a larger dataset of training examples we could create a model with high enough accuracy that medical professionals could comfortably rely on to calculate the BMI of their patients. This would allow them to gain insights into the patients health which would otherwise be too difficult or expensive to get. With larger amounts of data and an even more accurate model, this machine learning project could have large, positive impacts in the medical world. If doctors were to use the model and input the measurements of their patients, we could continue to improve our models by retraining with more training examples as the available data grows

8. Acknowledgements

We would like to acknowledge Fedesoriano for providing the data set on Kaggle[1], which allowed us to create and analyze our machine learning models. We would also like to thank Dr. A. Garth Fisher who originally gave permission to distribute this data set for non-commercial purposes. In addition we would like to acknowledge Professor Raschka for providing us with the tools we needed to successfully complete this project through his lectures and teaching.

9. Contributions

Model Creation:

- Hunter focused most on building the k-nearest-neighbors model.
- Rachel mostly worked on the decision tree model.
- Casey built the remainder of the models, with assistance from Rachel and Hunter in analyzing and interpreting the model results.

Analysis and Report:

- Casey primarily worked on the technical aspect of our project analysis, creating visualizations and charts in order to analyze our models.
- Hunter analyzed the model he worked on, and worked on the related work. He dove into the ways in which researchers in the past have analyzed and worked with similar data, ensuring that our project is useful and accurate in the medical field.
- Rachel analyzed the model she worked on, and focused on the background and motivation behind our project. She researched why this project can be useful in the medical field, and why an easier way to calculate BMI can have positive impacts on patients and medical professionals.

References

[1] Body fat prediction dataset. https://www.kaggle.com/fedesoriano/body-fat-prediction-dataset. Accessed: 2021-12-6.

[2] What is a healthy body fat percentage. https://tanita.eu/blog/healthy-body-fat-percentage/. Accessed: 2021-10-11.

[3] python - why do I get nan when using sklearn R2 function?, Oct. 2020.

[4] K. H. Collins, B. Sharif, C. Sanmartin, R. A. Reimer, W. Herzog, R. Chin, and D. A. Marshall. Association of body mass index (bmi) and percent body fat among bmi-defined non-obese middle-aged individuals: Insights from a population-based canadian sample. *Canadian Journal of Public Health / Revue Canadienne de Santé Publique*, 107(6):e520–e525, 2016.

[5] S. K. In and K. A. Kong. Body mass index and severity/fatality from coronavirus disease 2019: A nationwide epidemiological study in korea. *PLoS One*, 16(6), 06 2021. Copyright - © 2021 Kang, Kong. This is an open access article distributed under the terms of the Creative Commons Attribution License: http://creativecommons.org/licenses/by/4.0/ (the "License"), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2021-06-23; SubjectsTermNotLitGenreText - China; South Korea.

[6] C.-Y. Jung, H. Park, D. W. Kim, H. Lim, J. H. Chang, Y. J. Choi, S. W. Kim, and T. I. Chang. Association between Body Mass Index and Risk of Coronavirus Disease 2019 (COVID-19): A Nationwide Case-control Study in South Korea. *Clinical Infectious Diseases*, 73(7):e1855–e1862, 08 2020.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.