Predicting Human Body Fat Percentage with Certain Anthropometric Data

Ai-Shuan Lee, Yufan Zhang and Amber Ye
Mar. 16, 2025

Abstract

Obesity rates have increased dramatically over the past few years, making it an urgent issue that should be addressed. In the United States, it's estimated that over 60% of adults are overweight or obese(World Health Organization,2025). In this research, we aim to provide an alternative way to better estimate the body fat percentage and help people have a better understanding of their health situation. Through our findings in the "Body Fat Extended Dataset" on Kaggle, we realized that BMI can sometimes be inaccurate. After running through three different regression models, linear, random forest and support vector, we created the BodyFat Calculator (BDF) and compared it with BMI estimates and the real body fat data. Our results indicate that our BDF calculator is more accurate than BMI at predicting body fat, offering a better approach to assessing an individual's health.

Introduction

Obesity is an urgent issue that needs to be addressed, not just in the US but worldwide. The prevalence of obesity and the severity of it have increased drastically for younger adults (Hales, 2020). However, the word "obesity" is often used vaguely. Traditionally, we rely on body mass index (BMI) to predict the body fat percentage. However, studies have shown that the traditional method of predicting body fat percentage using BMI can be misleading and inaccurate (Woolcutt, 2018). This inaccuracy stems from BMI's narrow consideration of only height and weight. BMI does not differentiate between muscle mass and fat. This can lead to an inaccurate estimate of body fat percentage for people who have high and lean muscle mass (Frankenfield et al., 2001). Additionally, BMI fails to consider fat distribution. The distribution of body fat can significantly affect the risk of metabolic disease(Romero-Corral et al., 2010). In this study, we ran a correlation analysis through our data set to determine which anthropogenic factors contribute the most to the body fat percentage. We then use the three most correlated traits to run through regression models to create a better way of estimating body fat.

Literature Review

BMI is widely used but is considered inaccurate when estimating body fat by using people's height, weight, and sex. Woolcott and Bergman (2018) proposed relative fat mass(RFM), which is a better way to predict body fat using height and waist measurements. However, RFM needs validation across different populations. Nowadays, people care more about

personal health. We tried the well-known body fat calculator on the website, but it is not accurate according to our data. This leads us to think about making a body fat calculator by doing a correlation between body fat and different variables (hip circumference, knee circumference, weight, height, abdomen circumference…), which of them will be most correlated with body fat. When we get the correlation, we are able to choose the model we want to create for making a body fat calculator. Linear regression model is a great model to use in our datasets because the body circumferences are positively correlated with body fat, and it is kind of linear. However, non-linear regression models such as random forest and support vector machines could also be used to test the correlation between our variables and body fat.

Data Description

Our study utilized the "Body Fat Prediction Extended" dataset that is sourced from Kaggle. It comprises 16 anthropometric traits collected from 252 male and 184 female individuals. The dataset was originally compiled by Dr. A. Garth Fisher, who authorized its free distribution and non-commercial use. The original dataset only included male participants, hence for our study, we used the extended version which incorporates female participants. The recorded variables include body fat percentage, sex, age, weight(kg), height(m), neck, chest, abdomen, hip, thigh, knee, ankle, biceps, and forearm circumference(cm).

Exploratory Data Analysis & Visualization

The initial phase of data preprocessing involved assessing the dataset for missing values and outliers. While no missing values were detected, several outliers were identified, which could potentially distort the results. For a more accurate result, we used the Interquartile Range (IQR) method to systematically detect and remove them. Additionally, we replace the variables for "gender" from categorical to numerical. Male is represented by 1 and female by 0. We also dropped the "original" column from our data for simplicity purposes as it does not provide additional information for our study. As the dataset did not provide Body Mass Index (BMI) values, they were computed using the standard formula, "weight /(height^2)", where weight is measured in kilograms (kg) and height in meters (m). This calculation was necessary to facilitate comparisons between BMI and body fat percentage in later analyses.

Visualization

Body Mass Index (BMI) is widely used to estimate body fat percentage, but we wanted to test its accuracy in categorizing individuals as underweight, normal, or overweight. The standard BMI classification states that:

1. Underweight: BMI < 18.5 and body fat percentage < 14%
2. Overweight: BMI > 25 and body fat percentage > 25%

We applied K-clustering on BMI and BodyFat data to group individuals into three clusters. It then visualizes these clusters on a scatter plot, highlighting thresholds for underweight and overweight classifications. We found out that there are individuals who were categorized into the wrong classification if we just use BMI as the standard.

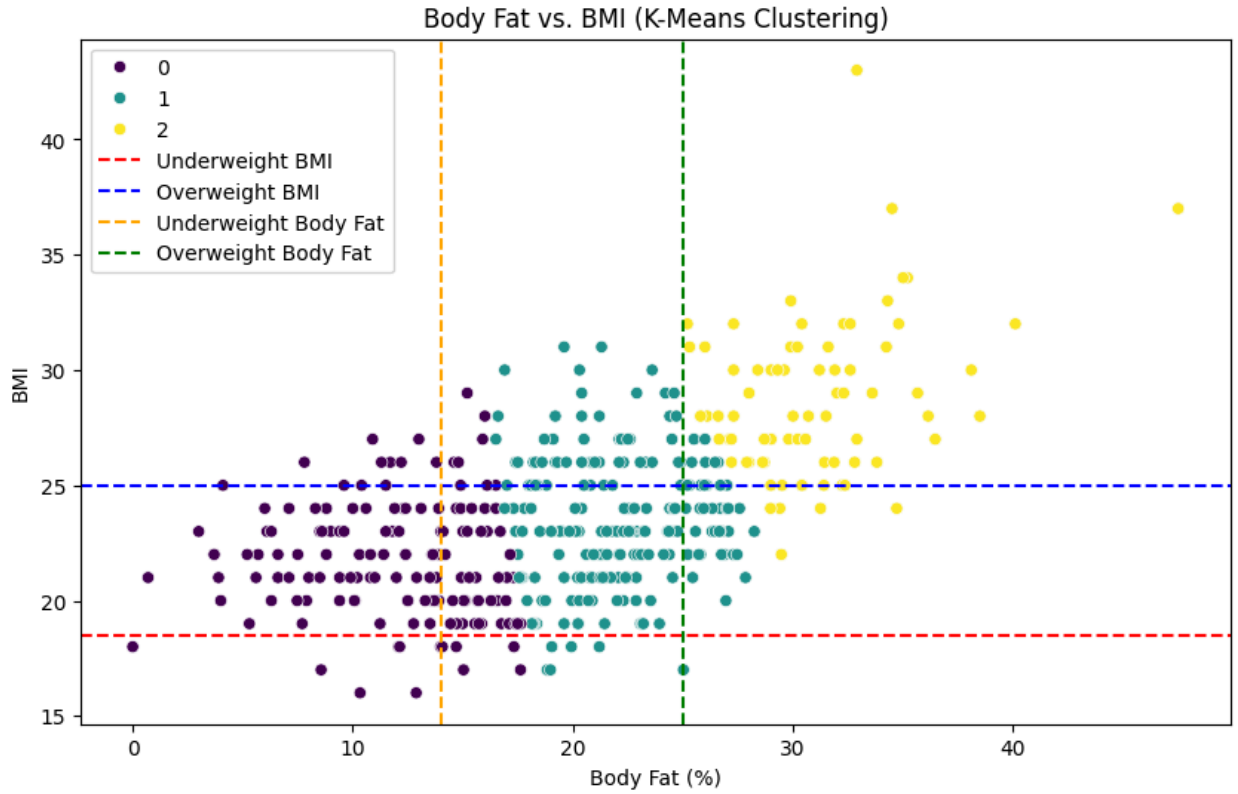

Fig.1 BodyFat Vs. BMI Scatterplot. The four lines indicate thresholds for body fat percentage and BMI regarding classification of underweight and overweight.

To evaluate the limitations of BMI as a predictor of body fat percentage, we conducted a correlation analysis between body fat percentage and all anthropometric traits in the dataset. A correlation heatmap (Fig. 2) was generated to visualize the interrelationships among these variables.

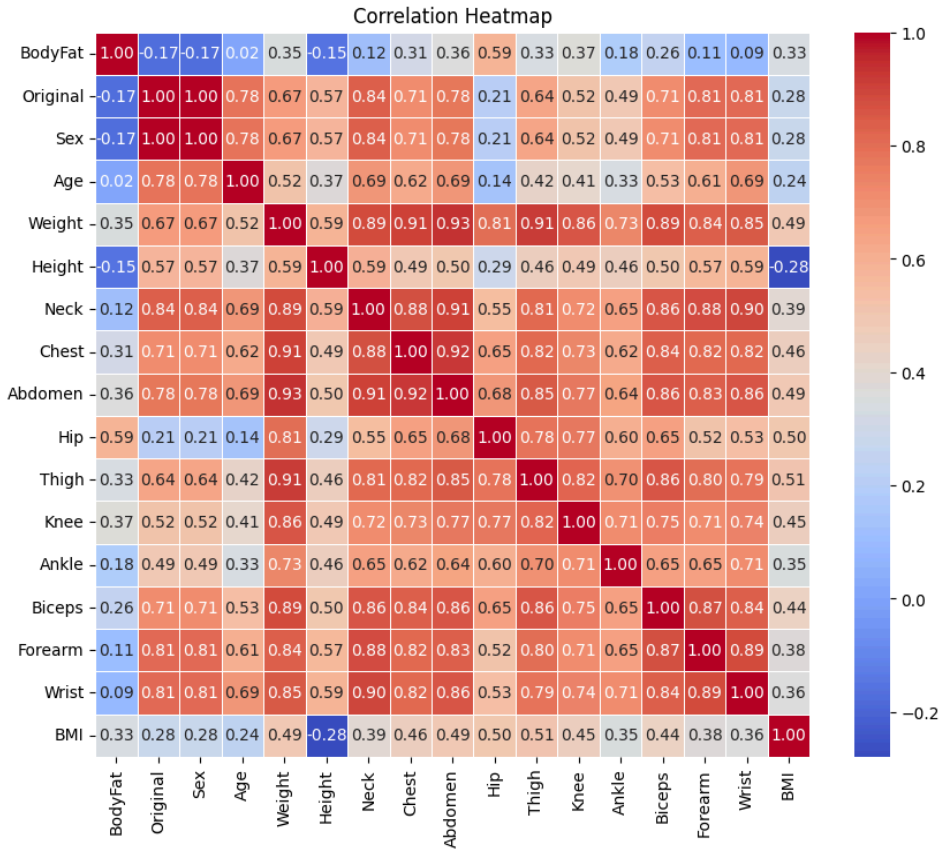

Fig.2 Correlation heatmap of each anthropometric trait. Blue indicates a weak correlation while red indicates a strong correlation. In this correlation heatmap, the standard variables for calculating BMI, weigh and height, are both showing weak correlation. T

Notably, contrary to the assumptions underlying BMI, the strongest correlations were not observed with sex, weight, or height. We further constructed a bar chart (Fig.3) illustrating the correlation coefficients of each trait with body fat percentage. The analysis revealed that the three most strongly correlated variables were the hip, knee and abdomen circumference.

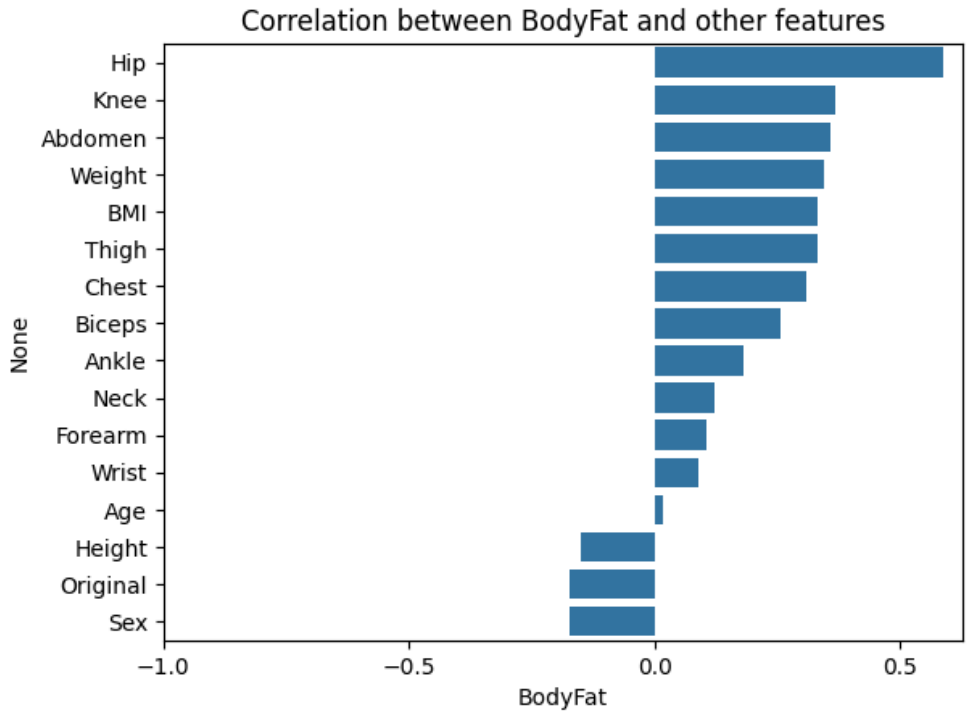

Fig.3 Correlations between body fat and other anthropometric traits using bar graphs. The top three most correlated variables were "hip", "knee" and "abdomen".

Proposed Methodology

We chose 6 variables: hip, knee, abdomen, weight, thigh, chest circumference to test three models we had chosen, which are Linear regression, Random forest, and Support vector. We will import those 6 variables as X=df.

After training each model, we will choose which model is best for producing a function for predicting body fat function by comparing their mean squared error (MSE) and r-squared (R^2).

Regression Analysis

These findings suggest that body fat percentage is more accurately predicted using specific anthropometric measurements rather than BMI alone. Using these three strongest predictors, we implemented a linear regression model in Google Colab to generate a prediction equation for body fat percentage. The resulting equation is $-41.28+0.93X-0.89Y+0.039Z$, where X, Y, Z each is the hip, knee, and abdomen circumference. The R^2 value for this model was 0.2977, indicating a moderate fit.

To improve model accuracy, we ran the regression again using the six most correlated variables, including hip, knee, abdomen, thigh, chest circumferences, and weight, which increased the R^2 value to 0.4341. Despite the improved fit in our expanded model, for the sake of simplicity and practicality, we decided to use the original three-variable equation as our final model.

Model Comparison & Results

Our study compared three different regression models on Google Colab to provide the most accurate prediction. The chosen models were simple linear regression, random forest regression, and support vector regression. The comparison is shown through a table below (Table 1).

Models	Linear Regression	Random Forest	Support Vector
MSE	32.4965	26.1378	34.4417
RMSE	5.7006	5.1125	5.8687
R^2	0.2977	0.4351	0.2557

Table 1. A comparison of the mean squared error (MSE), root mean squared error (RMSE) and r-squared (R^2) of the three regression models. The result indicates that the random forest regression model is able to generate the most accurate prediction.

The results indicate that the random forest model is able to provide the most accurate prediction for our dataset. With an MSE of 26.1378 and an R2 score of 0.4351. Which means about 43.51% of the variance in the target variable can be explained by the model. However, the process to get the prediction using the random forest model is too complicated (fig.4). For simplicity, our study will be using the function generated by the linear regression model, which provided a slightly better prediction than the support vector regression model (R^2: 0.2977 > 0.2557). We can know from R^2 the linear regression model performs great.

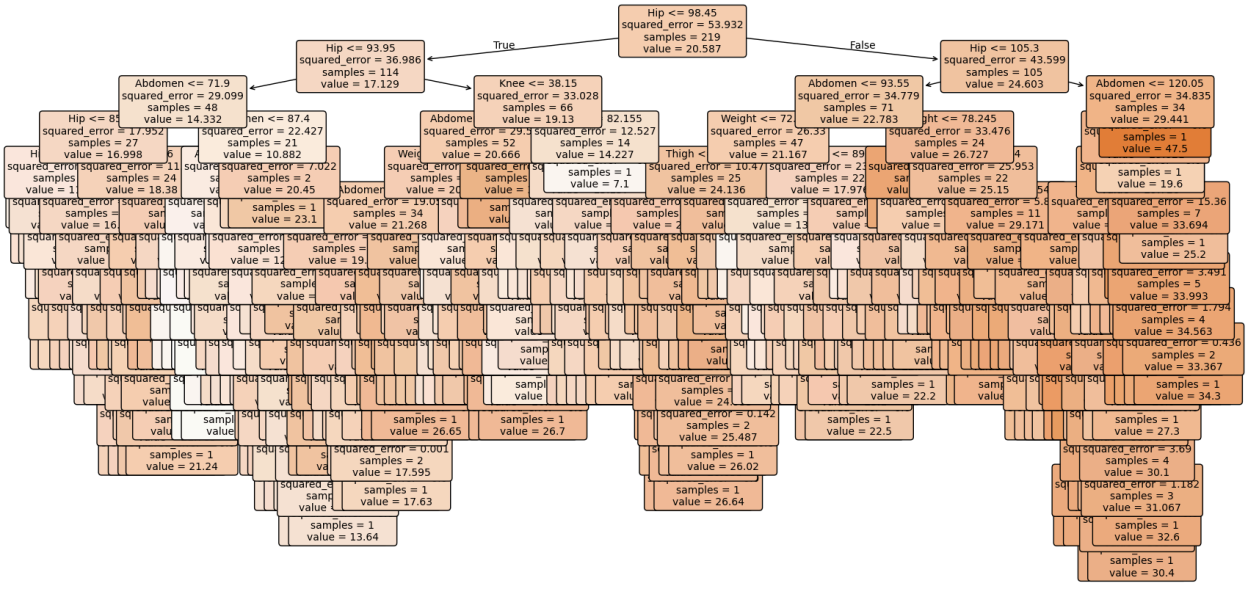

Fig.4 The decision trees that one must go through to get the prediction value for body fat. The decision trees are too complicated for further calculation.

Conclusion & Discussion

The study used multiple regression models to predict body fat percentage based on different body measurements from the dataset. After visualizing the data, the analysis indicated that the three body traits, hip, knee, and abdomen circumference, are the most correlated body traits with body fat percentage. Using these three variables, we ran the data through three regression models to compare the accuracy of the prediction model. Among the models tested, the random forest regression model was able to provide the most precise predictions. However, for simplicity, this study will be using the function generated by simple linear regression for further use. Although it is not the most accurate model for the dataset, it is able to provide an easier function for public use. Finally, the study developed a web app, called the Body Fat Calculator, which uses hip, knee, and abdomen circumference to estimate body fat percentage. Our calculator is more accurate than BMI and RFM in predicting body fat.

The findings also suggest that BMI is not a reliable predictor of body fat percentage, and it is often misleading individuals. Our model, based on hip, knee, and abdomen circumference, offers a more accurate and accessible method for estimating body fat. This can be used in health assessments and fitness tracking. The Body Fat Calculator also can be used in some mobile health apps. It also provides individuals with a personalized and data-driven approach to understanding their body composition. In the future, people can use Body Fat Calculator to estimate their body fat online, by using their measurements. This can help individuals better understand their body composition and make healthier decisions. Also, it is a useful health tracker. For further uses, it could be used in an exercise and diet tracking app. Finally, people would know what is affecting their body fat and how to gain more muscle and become healthier. It could be an effective tool related to our health.

Work Citation Page

"Data Page: Obesity in adults". Our World in Data (2025). Data adapted from World Health Organization. Retrieved from https://ourworldindata.org/grapher/share-of-adults-defined-as-obese [online resource]
https://www.kaggle.com/code/elvinrustam/bodyfat-prediction-regression-tutorial
https://www.kaggle.com/datasets/simonezappatini/body-fat-extended-dataset

Frankenfield, D. C., Rowe, W. A., Cooney, R. N., Smith, J. S., & Becker, D. (2001). Limits of body mass index to detect obesity and predict body composition. Nutrition (Burbank, Los Angeles County, Calif.), 17(1), 26–30. https://doi.org/10.1016/s0899-9007(00)00471-8

Hales, C. M., Carroll, M. D., Fryar, C. D., & Ogden, C. L. (2020). Prevalence of Obesity and Severe Obesity Among Adults: United States, 2017-2018. *NCHS data brief*, (360), 1–8.

Romero-Corral, A., Somers, V. K., Sierra-Johnson, J., Korenfeld, Y., Boarin, S., Korinek, J., Jensen, M. D., Parati, G., & Lopez-Jimenez, F. (2010). Normal weight obesity: a risk factor for cardiometabolic dysregulation and cardiovascular mortality. European heart journal, 31(6), 737–746. https://doi.org/10.1093/eurheartj/ehp487

Woolcott, O. O., & Bergman, R. N. (2018). Relative fat mass (RFM) as a new estimator of whole-body fat percentage: A cross-sectional study in American adult individuals. *Scientific Reports, 8*, 10980. https://doi.org/10.1038/s41598-018-29362-1

Woolcott, O. O., & Seuring, T. (2023). Temporal trends in obesity defined by the relative fat mass (RFM) index among adults in the United States from 1999 to 2020: a population-based study. *BMJ open*, *13*(8), e071295. https://doi.org/10.1136/bmjopen-2022-071295