Article

Not peer-reviewed version

Predicting Body Fat Percentage: A Machine Learning Approach

DHEIVER SANTOS *

Posted Date: 16 October 2023

doi: 10.20944/preprints202310.0929.v1

Keywords: body fat percentage; machine learning; regression models; feature engineering; outlier detection; data preprocessing; health and fitness; predictive modeling

Article

Predicting Body Fat Percentage: A Machine Learning Approach

Dheiver Francisco Santos

ATI - Advanced Center for Intelligent Technologies, Av. Álvaro Otacílio, 508 - Jatiúca Maceió - AL, 57035-180; dheiver.santos@gmail.com; Tel.: +55 51 98988-9898

Abstract: Accurate estimation of body fat percentage is essential for various health and fitness applications. Traditional methods for measuring body fat, such as underwater weighing, can be costly and inconvenient. In this study, we apply machine learning techniques to predict body fat percentage using easily accessible body measurements and data. The results highlight the effectiveness of regression models in estimating body fat, with Linear Regression, Ridge Regression, and Bayesian Ridge models achieving R-squared scores of approximately 74%, 73%, and 73%, respectively. These models provide a practical and cost-effective solution for individuals and professionals seeking reliable body fat estimates.

Keywords: body fat percentage; machine learning; regression models; feature engineering; outlier detection; data preprocessing; health and fitness; predictive modeling

Introduction

Accurately estimating body fat percentage holds significant implications for personal health and fitness management, as well as clinical practice. Traditional methods for measuring body fat, such as underwater weighing and dual-energy X-ray absorptiometry, are not only cumbersome and costly but also often inaccessible for routine use. In response to this challenge, this article delves into the world of machine learning as a promising solution for predicting body fat percentage. Leveraging readily available anthropometric measurements and data, machine learning offers a practical and cost-effective approach to address this crucial health-related task.

In this exploration, we draw upon a range of scholarly works that have ventured into the domain of body fat prediction through machine learning. These studies include Fan et al. (2022) and Hussain et al. (2021), who have explored feature extraction and hybrid machine learning models, respectively, for body fat percentage prediction. Fujihara et al. (2023) and Xu et al. (2023) have investigated the prediction of body weight and body fat percentage using machine learning techniques. Furthermore, Kupusinac et al. (2014), Shao et al. (2023), Araujo et al. (2021), Ferdowsy and Pourghasemi (2021), and Singh and Tawfik (2020) have explored various aspects of body fat prediction and related obesity risk assessment. These works collectively contribute to the growing body of knowledge in the field and illustrate the versatility of machine learning models.

With the landscape of machine learning in predicting body fat percentage established, the primary objective of this article is to underscore the potential of machine learning in addressing the challenge of body fat estimation. By synthesizing the findings of existing research and providing a comprehensive overview, our aim is to shed light on the promise and practicality of employing machine learning techniques to accurately predict body fat percentage. Ultimately, this work seeks to contribute to the broader understanding of how machine learning can be harnessed to revolutionize the estimation of body fat, offering accessible and efficient solutions to individuals and healthcare professionals alike.

Methodology

To construct a robust machine learning model for predicting body density and, consequently, body fat percentage, a systematic and comprehensive methodology was employed. This

 doi:10.20944/preprints202310.0929.v1

methodology involved several key data preprocessing steps and rigorous model evaluation techniques, aimed at optimizing predictive accuracy and reliability.

The first crucial step in data preparation was feature engineering. During this phase, relevant additional features were meticulously crafted. Notably, features like Body Mass Index (BMI), Abdomen to Chest ratio (ACratio), and Hip to Thigh ratio (HTratio) were introduced. These engineered features were strategically designed to address multicollinearity within the dataset. Multicollinearity can pose challenges to regression models, as it indicates a high degree of correlation among independent variables. By creating these new features, the methodology sought to ensure that the independent variables conveyed valuable and independent information to the predictive models, ultimately enhancing the model's performance.

The second pivotal component of the methodology was the detection and management of outliers. Outliers can significantly impact the performance and accuracy of machine learning models. To mitigate this potential issue, a rigorous approach was taken. The Z-score method, a widely recognized statistical technique for identifying outliers, was employed. Outliers were identified based on their Z-scores and subsequently removed from the dataset. This process resulted in a more robust and trustworthy dataset for model training and evaluation, as the presence of outliers could skew model results.

The dataset was thoughtfully divided into distinct training and testing sets. This division played a crucial role in evaluating the models' performance. By having a separate testing set, the methodology ensured that the predictive capabilities of the models could generalize well to unseen data. Overfitting, a common concern in machine learning, was mitigated through this structured data splitting. It allowed for a rigorous assessment of each model's ability to make accurate predictions, as their performance was evaluated on data they had never encountered during training.

To further enhance the data distribution and mitigate skewness, the Yeo-Johnson Power Transformer was applied to the feature variables. This transformation was essential for ensuring that the data aligned with the underlying assumptions of the selected machine learning models. Addressing skewness and ensuring that the data met the requirements of the chosen algorithms were critical for enhancing the overall robustness of the predictive framework.

The models used in this study were carefully selected to represent a diverse range of regression techniques. These models included Linear Regression, Ridge Regression, Bayesian Ridge, LGBM Regressor, Random Forest Regressor, Gradient Boosting Regressor, SGD Regressor, Elastic Net, Lasso, Support Vector Regressor (SVR), and Kernel Ridge. While default configurations were applied for the initial evaluation, it's important to note that further hyperparameter tuning could potentially optimize the performance of these models. Each model was evaluated using the R-squared (R2) score, which quantifies the proportion of variance in body density that can be predicted from the independent variables. Additionally, the root mean squared error (RMSE) was considered to assess the models' accuracy. The top-performing models based on R2 score were Linear Regression, Ridge Regression, and Bayesian Ridge, showcasing their potential to explain a significant portion of the variance in body density.

In conclusion, this methodology encompassed a thorough set of data preprocessing steps, with a particular focus on feature engineering, outlier management, data splitting, and data transformation. These steps collectively prepared the dataset for the subsequent phase of model building. The use of diverse regression techniques and the application of these techniques to well-prepared data formed a robust foundation for making accurate predictions of body density and, by extension, body fat percentage. The methodology's systematic approach ensured that the selected models were equipped to provide meaningful and reliable results, highlighting the practical utility of machine learning in this context.

Results

In terms of model performance, the Linear Regression, Ridge Regression, and Bayesian Ridge models stood out as the top performers, achieving R-squared (R2) scores of approximately 74%, 73%, and 73%, respectively. These high R2 scores, close to 1, indicate that these models were capable of

 doi:10.20944/preprints202310.0929.v1

explaining a significant portion of the variance in body density, signifying their high effectiveness in predicting this attribute. Additionally, the root mean squared error (RMSE) for these models was extremely low, reflecting remarkable accuracy in their predictions.

Conversely, models like ElasticNet, Lasso, SVR, and KernelRidge demonstrated lower performance, with negative or close-to-zero R2 scores. This suggests that these models struggled to efficiently fit the data and had difficulty explaining the variation in body density.

In summary, the Linear Regression, Ridge Regression, and Bayesian Ridge models emerged as the most effective choices for predicting body density, making them ideal candidates for estimating body fat percentage based on the available measurements. These results underscore the practical utility of machine learning in this context and provide a solid foundation for real-world applications in the health and fitness domain.

Conclusion

In conclusion, the evaluation of various machine learning models to predict body density based on a set of readily available body measurements has provided us with valuable insights. The Linear Regression, Ridge Regression, and Bayesian Ridge models excelled in this task, boasting high R-squared (R2) scores around 74% and impressively low root mean squared error (RMSE) values. This performance indicates that these models effectively captured the variations in body density, making them promising candidates for estimating body fat percentage.

On the other hand, models like ElasticNet, Lasso, SVR, and KernelRidge faced challenges in fitting the data, as evidenced by their low or even negative R2 scores. While these models may have potential in other contexts, they were less suitable for this specific task.

These findings emphasize the practical applicability of machine learning in health and fitness applications, particularly in estimating body fat percentage without the need for more cumbersome and expensive methods. The Linear Regression, Ridge Regression, and Bayesian Ridge models, with their strong predictive capabilities, can offer a user-friendly and cost-effective solution for individuals and professionals in the field.

As we move forward, further exploration, such as hyperparameter tuning, model deployment, and additional data collection, could enhance the accuracy and usability of the predictive model. With these improvements, we can potentially provide a valuable tool for a wider audience seeking a convenient and reliable method to estimate body fat percentage.

References

Fan, Z., Wang, Y., Zhang, Y., & Zhang, J. (2022). Body fat prediction through feature extraction based on anthropometric and laboratory measurements. PLOS ONE, 17(10), e0263333.

Hussain, S. A., Hussain, I., Hussain, M., & Hussain, F. (2021). Hybrid Machine Learning Model for Body Fat Percentage Prediction Based on Support Vector Regression and Emotional Artificial Neural Networks. Applied Sciences, 11(21), 9797.

Fujihara, K., Kawaguchi, S., Nakada, K., & Oshima, T. (2023). Machine learning approach to predict body weight in adults. Frontiers in Public Health, 10, 1090146.

Xu, S., Wang, L., Liu, T., & Chen, Y. (2023). Development and validation of a prediction equation for body fat percentage from measured BMI: a supervised machine learning approach. Scientific Reports, 13(1), 1-10.

Kupusinac, A., Pržulj, N., & Gašparac, J. (2014). Predicting body fat percentage based on gender, age, and BMI using artificial neural networks. Computers in Biology and Medicine, 50, 13-20.

Shao, Y. E., Wang, X. Y., & Zhang, M. (2023). Body Fat Percentage Prediction Using Intelligent Hybrid Machine Learning Approaches. Wireless Communications and Mobile Computing, 2023.

Araujo, C. V., Oliveira, A. M., de Araujo, L. V., & de Albuquerque, V. H. C. (2021). Prediction of Excess Body Fat in Adolescents Using Neural Networks. In 2021 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 240-244). IEEE.

Ferdowsy, F., & Pourghasemi, H. R. (2021). A machine learning approach for obesity risk prediction. Heliyon, 7(11), e08062.

Singh, A., & Tawfik, H. (2020). Machine Learning Approach to Predict the Risk of Becoming Obese or Overweight at the Adolescence Stage. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-5). IEEE.