2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) | 979-8-3503-1590-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACCAI58221.2023.10200647

Body Fat Prediction using Various Regression Techniques

Nikhil Mahesh[1], Peeta Basa Pati[2], K. Deepa[3], Suresh Yanan[4]

[1,2]*Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Bengaluru, India*
[3]*Department of Electrical and Electronics Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru,*
[4]*Department of Engineering, University of Technology and Applied Sciences-Al Musannah Sultanate of Oman*

E-mail: nikhilmahesh89@gmail.com, bp_peeta@blr.amrita.edu, k_deepa@blr.amrita.edu,suresh@act.edu.com

Abstract- Predicting body fat percentage is essential for addressing the obesity problem. This paper compares the performance of several machine learning models based on Regression, to predict the body fat percentage. Using a dataset of 252 participants with information on age, weight, height, and fat percentage, the models were assessed based on multiple performance criteria, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Squared Error(MSE). The results demonstrates that Random Forest Regressor surpass other models with a lower RMSE of 0.276. These findings suggest that machine learning models can be a valuable tool for precise BFP, the use of machine learning provides a faster and more precise method for predicting body fat percentage. Overall, the study's results suggest that machine learning models can be valuable tool for accurate body fat percentage prediction.

Keywords: Random Forest, Body Fat Prediction, Decision Tree, Lasso Regression, LGBM, Linear Regression, and Ridge Regression.

I. INTRODUCTION

In the public health sector, obesity is significant public health problem that can lead to a range of serious medical conditions, which can lead to a number of diseases like heart disease, diabetes, cancer, and musculoskeletal ailments. The BFP must be kept in excellent condition for human health. It is difficult to find practical and accurate methods for calculating body fat, nevertheless [1]. For instance, it was claimed that hydrostatic weighing was a valid step for determining the amount of body fat, but it is not practical. Thus, it is crucial to be able to precisely forecast the body fat percentage in the machine learning model. ML models were used to predict the body fat percentage.

Zongwen, Raymond & Zhongi [1] investigated the use of feature extraction technique based on laboratory measurements and anthropometric for predicting body fat percentage. The feature extraction techniques used in the study included linear discriminant analysis (LDA) and principal component analysis (PCA) which were used to reduce the dimensionality of the feature space and identifies the most important features for prediction and used ML algorithms like linear regression, support vector regression (SVR) and random forest regression to predict based on extracted features. The authors suggested that their framework could be valuable tool for management of obesity-related diseases.

Zachary, April and Cham [2] have utilized various techniques for measuring body fat, such as X-ray absorptiometry(DXA), bioelectrical impedance analysis (BIA) and skinfold thickness measurements. They had conducted a cross-sectional and longitudinal studies to develop and validate their models and compared the results to existing models and reference methods for body fat measurement. Their findings suggested that certain predictor variables such as body mass index (BMI) and waist circumference are strong indicators of body fat can be used to develop accurate and reliable prediction models. They had highlighted the importance of developing robust prediction models for body fat in American adults. These models have potential to improve clinical assessments of body composition more effective interventions for obesity and related health conditions.

This paper compares the accuracy of various regression techniques, such as linear regression, ridge and lasso regression, decision tree and random forest repressor, gradient boosting, LGBM, and single gradient, in predicting body fat percentage, it is based on the root mean square error (RMSE), mean square

error (MSE) and mean absolute error (MAE) criteria. This comparison will help determine which model is the most accurate in predicting body fat percentage.

II. Dataset Description

To construct a BFP prediction model that can predict BFP with high accuracy rates using the pertinent anthropometric features and identify the most important elements that influence the prediction model, primary information was gathered for this project. The dataset consists of 15 attributes (i.e. Age, Height, Weight, Ankle, Biceps, Forearm, Wrist, Chest, Neck, Hip, Abdomen, thigh, knee, and body fat), the dataset consists of 252 records. The obtained data only included a single observation for each individual and did not include time-series data that might be used to forecast the BFP in the future based on an individual's most recent anthropometric and laboratory measurements. To reduce the time and expense required to get BFP, the study concentrated on the precise prediction of a person's recent BFP. BFP is denoted as Y, using Siri's equation(1).

$$Y = \frac{495}{D} - 450 \qquad (1)$$

The given dataset [16] have density, weight (lbs.), body fat, hip (cm), neck (cm), abdomen (cm), thigh (cm), ankle (cm), biceps (cm), thigh (cm), forearm(cm), wrist (cm), height (inches), chest (cm), age (in years). There was no null values present in the data. Checking the correlation in the dataset between various features, we can see in Figure 1, abdomen, chest, hip and limbs are highly correlated to each other.

In figure 2, it shows that outliers was present in arms, biceps, forearm, thighs, ankle, hip, abdomen, chest, abdomen, density, body fat, and density variables. We decided to remove the outliers in this study because they significantly impacted the mean and median of the data, which would have led to incorrect conclusions about the data, therefore it was necessary to remove the outliers. Figure 3 shows the boxplot of various features after the removal of outliers was done.

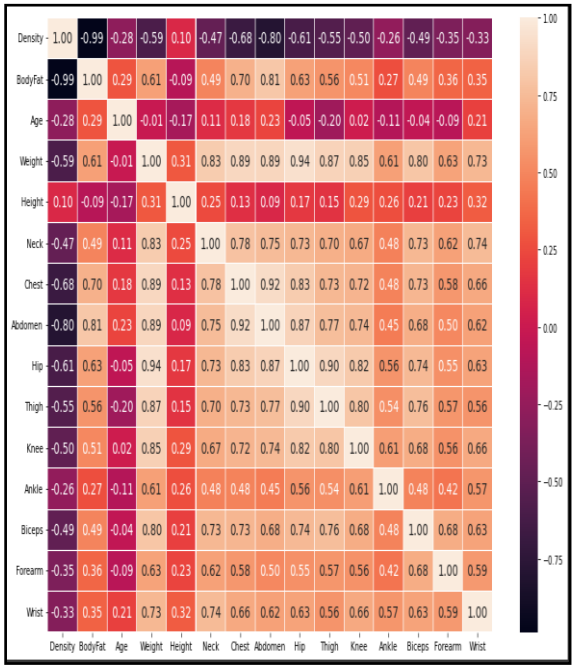

Figure 1: Heatmap of Correlations between different features

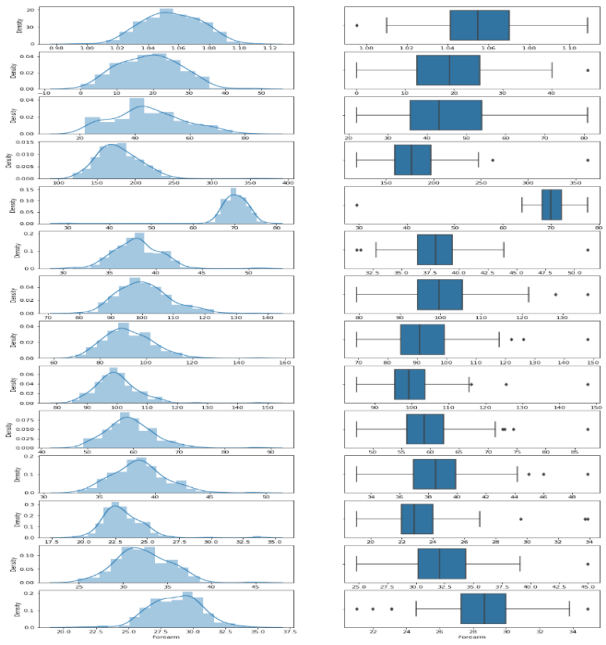

Figure 2: Presence of Outliers

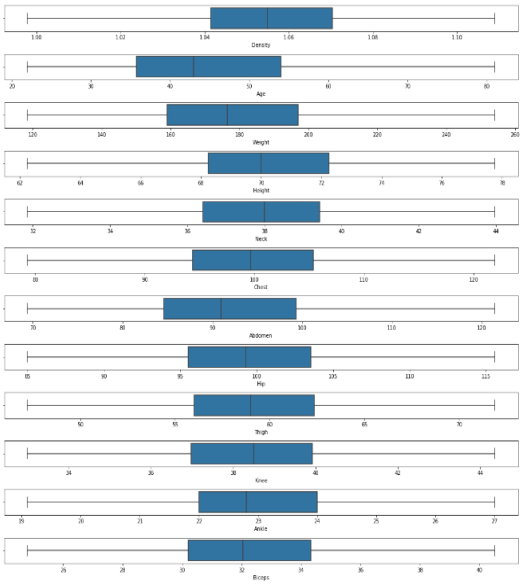

Figure 3: Removal of outliers

III. Machine Learning Models

A. LASSO Regression

The operator with the least selection and absolute shrinkage, or a kind of linear regression are LASSO with a method for reducing the regression coefficient of the predictor variable. In comparison to normal linear regression, LASSO regression is more suited for managing data with multicollinearity issues in a particular prediction model [11].

B. LR(Linear Regression)

It's a supervised learning technique that produces continuous value outputs and predictions. The outcome will be looking at the line of perfect fit that can reduce a squared difference between the predicted and observed values. The line, which depicts the connection between the dependent and independent variables, is used to make predictions about the dependent values that are based on updated values of the independent values [13].

C. GB (Gradient Boosting)

For classification and regression issues, the ensemble machine learning technique gradient boosting is utilized. It creates a powerful model that can accurately forecast the future by combining several weak models. Iteratively training weak models and changing their weights in a way that corrects the errors of earlier models is how the algorithm operates [4].

The gradient boosting approach generates a new model at each iteration and fits it to the loss function's negative gradient. The combined forecasts of all weak models result in the final prediction[12].

D. Regression Tree

It is a specific type of ML technique used in classifying the input values. Each internal node represents the “test” while the "result" by each branch, and the class label or prediction by each leaf node. This creates a tree-like model of deciding and their outcomes. Maximizing information gained at each split will result in a decision boundary that clearly and visibly divides the classes. Decision trees can handle both category and numerical data and are quick to train [5].

E. Ridge Regression

In multiple-regression models with closely related independent variables, ridge regression is a technique for estimating the coefficients of all regression machine learning models [9].

F. RFR (Random Forests Regression)

A learning method and numerous decision trees are the foundation of the supervised learning system known as Random Forest, the decision Trees do not interact and all computations are performed simultaneously as they are built in Random Forest since it employs a bagging strategy. With RF, challenges involving classification and regression may both be resolved [14].

G. Single Gradient

The best answers for many issues could be discovered using the all-purpose optimization approach known as gradient descent. The basic concept is to iteratively alter parameters to minimize the cost function. The step size, which is defined by the learning rate hyper parameters, is a crucial Gradient Descent (GD) parameter [7]. The method will need to go through multiple rounds before it converges, which will take a long time. If the learning rate is too high, we run the risk of going above the ideal value.

H. LGBM

LGBM, a gradient-boosting ensemble method based on decision trees, is used by the Train Using AutoML tool. A decision tree-based method called LightGBM may be used to solve both classification and regression issues. LightGBM has been specifically created for outstanding performance with scattered systems [6]. LightGBM builds decision trees that

evolve leaf-wise, meaning that, given a condition, only one leaf may be split for each tree, depending on the benefit. Leaf-wise trees might overfit at times, especially with smaller datasets [15]. By restricting the depth of the tree, overfitting may be avoided.

IV. Performance Measures

A. RMSE

In machine learning, the RMSE (Root Mean Square Error) is a popular assessment statistic for regression issues. Root of the squared difference between predicted and real values are used to compute the RMSE. A lower RMS error value indicates that the model is better suited to the data [3].

B. MSE

A typical evaluation criterion for regression concerns in machine learning is MSE (Mean Square Error). Mean square of anticipated and actual value differences determines the MSE. An improved fit of the model to the data is shown by a lower MSE value. The root over the MS error is all that the RMS error [8].

C. MAE

A typical evaluation tool for regression problems in machine learning means absolute error (MAE). The mean between expected and actual value is used for determining the MAE. MAE is a more reliable statistic for assessing model performance than MSE since it is not subject to outliers [1, 8].

V. Results

Scores in various machine learning techniques were compared, by checking their RMSE , MSE, and MAE score, from table 1, RFR had a low RMSE Score and DTR had a high RMSE score. After Cross-validation was done, in Table-1 DTR has a low RMSE score and DTR has a high RMSE score.

TABLE 1: SCORES FOR DIFFERENT MODELS

Model	RMSE	MSE	MAE
RF	**0.276**	**0.076**	**0.177**
XGB	0.293	0.086	0.210
LR	0.376	0.142	0.241
Ridge	0.537	0.289	0.431
XGB	0.551	0.304	0.442
Lasso	0.632	0.400	0.546
SGD	0.861	0.741	0.540
LGBM	0.898	0.807	**0.723**
DTR	**1.329**	**1.766**	0.496

From Figure 4, RF has a low RMSE of 0.276, whereas DTR has high RMSE score 1.329. To achieve a perfect RMSE score, a very low score is considered good. For MSE score, DTR (i.e Regression Tree) has a high MSE score of 1.766 and Random Forest has a low MSE, a perfect MSE score would be zero or lower score is better and inidcates a better fit between the predicted values and actual values. In MAE score, RFR has a low MAE score 0.177, whereas LGBM has a high MAE score of 0.723, a lower MAE score is considered better and indicates better performing model.

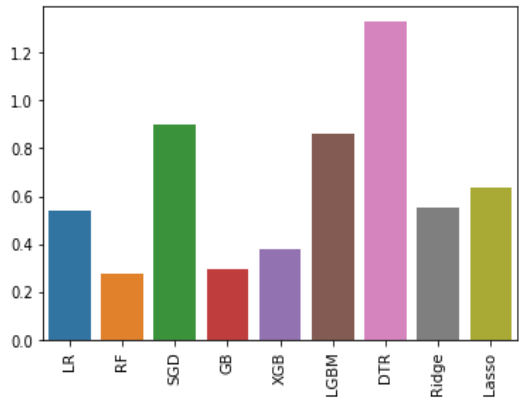

Figure 4: Different Regression models for RMSE Scores

TABLE 2: AFTER CROSS VALIDATION OF DIFFERENT ML MODELS

Model	RMSE	MSE	MAE
LR	1.117	2.031	0.574
LGBM	1.169	2.020	0.566
GB	1.183	2.019	0.576
XGB	1.188	2.005	0.560
DTR	1.196	1.993	0.552
Ridge	1.209	1.978	0.550
Lasso	1.211	2.007	0.562
SGD	1.211	2.010	0.571
RF	1.225	2.035	0.579

In Table 2 cross validation for different machine learning models was done to check for RMSE score, MAE score, MSE score and to conclude which one is better in different machine learning model.

In figure 5, it shows the plot for after cross-validation LR has a low RMSE score, whereas RF has a high RMS Error score. To achieve a perfect RMSE score, a very low score is considered good. To check for MSE after the cross-validation, RF has a high MSE score and DTR has a low MSE score, a perfect MSE

score would be zero or lower score is better and inidcates a better fit between the predicted values and actual values.

To check for MAE score after cross validation for MAE score, Ridge has a low MAE score and RF has a high MAE score, a lower MAE score is considered better and indicates better-performing model. Plotting for Cross Validation for different machine learning techniques

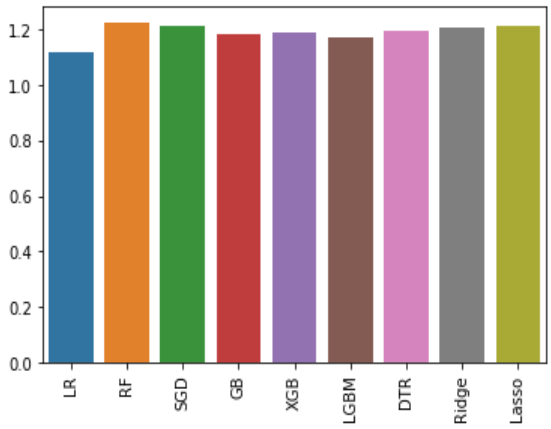

Figure 5: After cross-validation of different Regression models for RMSE Scores

VI. Conclusion

This study aimed for comparing different machine learning models in predicting body fat percentage. By analyzing that several factors such as Age, weight, height, density, and waist circumference have a significant influence on body fat percentage. Accurate prediction of body fat percentage using machine learning model has important factor for addressing obesity problems. In this study, a key dataset for a comparative study of body fat percentage was done and employed several regression techniques, including Lasso, Ridge, Linear, Random Forest, XGBoost, Gradient Boosting, LGBM, and Decision Tree. From table 1, RF (Random Forest) was having low RMS Error, MS Error, and MA Error scores,whereas DTR was having high RMS Error, MS Error, and MA Error scores. For RMS Error score, RF is 0.276 and DTR (Regression Tree) is 1.329. From table 2 after cross-validation, LR was having low RMS Error and MA Error scores, Lasso was having low MS Error and LGBM was having high RMS Error, MS Error, and MA Error scores. Best results in BFP can be obtained by adding more feature like lifestyle and environment factors in dataset. It can help in getting more accurate results or by using more machine learning techniques in order to get the best model for the comparative study.

References

[1] Fan Z, Chiong R, Hu Z, Keivanian F, Chiong F Body fat prediction through feature extraction based on anthropometric and laboratory measurements (2022) PLOS ONE 17(2): e0263333.

[2] Merrill Z, Chambers A, Cham R ("Development and validation of body fat prediction models in American adults") Obes Sci Pract. 2020 Jan 15;6(2):189-195.

[3] Devagopal AM, Ashwin V, Vishal Menon, "Prediction of Water Quality Parameters of River Periyar Using Regression Models," 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, Greater Noida, India, 2022, pp. 53-57.

[4] P. Ganesh, H. V. Vasu, and D. Vinod, "Estimation of Rainfall Quantity using Hybrid Ensemble Regression," 2019 9th International Conference on Advances in Computing & Communication, India, 2019, pp. 300-309.

[5] G. V. Sajan, P. Kumar, "Forecasting and Analysis of Train Delays and Impact of Weather Data using Machine Learning," 12th Int. Conf. on Computing Comm. & Networking Technologies, India, 2021, pp. 1-8,

[6] K. C. Kumar and Rajesh M, "Ethereum and Binance Price Forecasting Using Machine Learning," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-8,

[7] A. Ashok and C. P. Prathibhamol, "Improved Analysis of Stock Market Prediction: (ARIMA-LSTM-SMP)," 2021 4th Biennial Int. Conference on Nascent Technologies in Engineering (ICNTE), India, 2021, pp. 1-5,

[8] T. Aravind, S. Avinash and Jeyakumar. G., "A Comparative Study on Machine Learning Algorithms for Predicting the Placement Information of Under Graduate Students," 3rd International conference on IoT in Social, Mobile, Analytics & Cloud, India, 2019, pp. 542-546,

[9] S. Sonu, A. Suyampulingam, "Linear Regression Based Air Quality Data Analysis & Prediction using Python," IEEE Madras Section Conf., India, 2021, pp. 1-7.

[10] Kajal Rai, "Students Placement Prediction Using Machine Learning Algorithms", 2022

[11] D. Satish Kumar, et.al, "Predicting Student's Campus Placement Probability using Binary Logistic Regression", Int. J. of Innovative Technology and Exploring Engg, Vol. 8, No. 9, pp. 2278-3075, 2019.

[12] J. E. Ball and K. C. Luk, "Modeling spatial variability of rainfall over a catchment", J. Hydrologic Eng., vol. 3, no. 2, pp. 122-130, Apr. 1998.

[13] A. Parmar, K. Mistree, and M. Sompura, "Machine learning techniques for rainfall prediction: A review," in Proc. 4th Int. Conf. Innov. Inf. Embedded Commun. Syst. (ICIIECS), Mar. 2017, pp. 152–162.

[14] M. Lee, N. Kang, H. Joo, H. Kim, S. Kim, and J. Lee, "Hydrological modeling approach using radar-rainfall ensemble and multi-runoff-model blending technique," Water, vol. 11, no. 4, pp. 1–18, 2019.

[15] C. Frei and F. A. Isotta, "Ensemble spatial precipitation analysis from rain gauge data: Methodology and application in the European alps," J. Geophys. Res., Atmos., vol. 124, no. 11, pp. 5757–5778, Jun. 2019.

[16] Fedesoriano, Body Fat Prediction Dataset– https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset (2021).