International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 06, 2020
ISSN: 1475-7192

MACHINE LEARNING APPROACH FOR PREDICTING BODY FAT

[1]Vairachilai S, [2]Shubhangi V Urkude, [3]Gnanajeyaraman R

ABSTRACT-- *A human body needs a certain amount of fat to function properly. Fat controls the body temperature protects the organs, and store energy for body functioning. For the human body, it is important to assess the current status of the body fat in order to make correct decisions to improve health. Recently, the Machine Learning approach has empowered strong and accurate predictions on many healthcare applications. Regression analysis is one such supervised machine learning approaches which is used to analyze the significant factors which will affect the body depending on the fat content. In this paper, regression analysis techniques such as Multiple Linear Regression (MLR), and Support Vector Regression (SVR) are applied and analyzed to predict the body fat. The performance of the algorithms has been evaluated based on regression models validation metrics such as Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE).*

Keywords-- *Body Fat, Multiple Linear Regression, Support Vector Regression, Machine Learning.*

I. INTRODUCTION

Fat is important for the human body. Though, excess fat leads to obesity. Shiva Shanth Reddy Ainala et al., [1] proposed the regression model to predict body fat based on anthropometric variables such as thigh, knee age, weight, and gender, etc. J.P. Verma et al [2] proposed the regression model to predict the body fat. J.P Verma's model was used to assess the body fat content in college male's students within the age range of 18 to 24 years. St-Onge MP [3] studied the impact of aging on different body composition, body functioning, and body metabolism. Regression analysis is performed on the collected data. The result shows that muscle mass is reduced and fat mass is increased in older adults. In addition, intramyocellular lipid, and liver fat increased and changed the metabolic risk and physical function when compared to the younger adults. Martarelli D et. al. [4] proposed a regression analysis technique to find the body composition based on Body Mass Index (BMI), age, and body structure. The regression analysis is performed on male and female samples of the Italian population separately to estimate the body composition. Based on bioelectrical impedance and anthropometric analysis, the data is analysed and compared with observed masses with the help of Bland and Altman plots. The prediction equation is proposed to estimate body fat (Gómez-Ambrosi J et. al. [5]). The experiment is carried out by air displacement Plethysmography (ADP) method. The proposed equation is tested with 6510 samples which consist of 67% female and 33% male of the age group range from 18-80 years. The proposed equation provided more accurate result than

[1] Department of Computer Science and Engineering, Faculty of Science and TechnologyThe ICFAI Foundation for Higher Education (IFHE), Hyderabad – 501203, Telangana, India, vairachilai@ifheindia.org

[2] Department of Computer Science and Engineering, Faculty of Science and TechnologyThe ICFAI Foundation for Higher Education (IFHE), Hyderabad – 501203, Telangana, India, ushubhu@ifheindia.org,r.

[3] Department of Computer Science and Engineering, SBM College of Engineering and Technology Dindigul, Tamil Nadu 624005, India, gnanajeyaraman@gmail.com

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

the traditional method. Jackson et. al. [6] proved that the percentage of body fat is independent of age and gender for both men and women using polynomial regression. According to them, the percentage of body fat and BMI is proportional to each other. This model showed that the percentage of body fat for women depends upon age, race, and race-by-BMI, for men it is a function of age and age-by-BMI. Burton BT et. al. [7] summarized the findings of the 19 experts in the relevant area (Body fat). It contains the summary table for BMI values and weights. It shows the dependence of BMI on height & weight and provides the standard table for BMI values for clinical use. Chathuranga Ranasinghe et. al. [8] applied the regression analysis to find the body fat percentage. They divided the age group into three categories such as men & women as young, middle-aged and elderly. The result showed that gender, age, and BMI have nonlinear relationships whereas the percentage of body fat has linear relationships. Ifeoma F. Odo et. al. [9] explained how body composition will influence body fat. According to his findings, the percentage of body fat and BMI has a positive relationship with age in all aspects and negative relationship to body muscle and body water. The analysis is done on a sample of 780 people aged between 10-20 in two different regions of Enugu state of Nigeria. He concluded that BMI is used as a measure for the percentage of body fat. Pilar Fuster-Parra [10] proposed the simplest model using only BMI and Body Adiposity Index (BAI) to predict the percentage of body fat. Manuel Ramirez-Zea et al. [14] proposed a prediction equation to predict body fat using anthropometric measurements such as skinfold thickness, arm circumferences, and waist circumferences, etc. This study has been conducted in both rural and urban Guatemalan adults. This linear regression model predicted the percentage of body fat with reasonable accuracy. Su-Hsin Chang et. al. [11] reviewed body fat distribution in adults. They performed 17 different studies on a new technique called mortality analysis. The study showed that BMI is not an appropriate predictor for body fat because it will not consider the fat distribution with respect to age. They used three ranges of mortality such as the highest, mildest and lowest mortality. The result shows that the low mortality adult was overweight (mildly obese) & obesity and body fat are the risk factors for increased mortality in the population. Manish S. Bharadwaj et. al. [12] explained the relationship between various body composition factors such as whole body, body fat distribution, the sensitivity level of insulin, thigh composition with electron transport chain to drive respiration, mitochondrial content, and skeletal muscle. The experiment is performed on 25 samples including thirteen men and twelve women under the age of 65 years having BMI range 18-35kg/m^2. The result showed that skeletal muscle mitochondrial content, electron chain function, adiposity, and body composition are interrelated in normal adults. Bandana Sen et al. [13] developed an equation to predict body fat percentage in Indian infants and young children based on attributes such as age, thickness, and mid-arm circumference. In this paper, the body fat percentage is calculated using the dilution method and prediction equations. This result shows that the equation predicts the percentage of fat in children's growth in their early stages of life. C. Pongchaiyakul et al. [15] proposed a method to predict body fat using Anthropometric Measurements such as height, weight, and hip circumference, etc. This study has experimented on rural Thailand people belonging to the age group of 20 to 84 years old. This research work could have yielded better results in clinical research without the DXA (Dual-energy X-ray Absorptiometry) settings.

The rest of the paper is organized as follows: Section 2, presents a detailed description of the body fat dataset and pre-processing the dataset for the current study. Section 3, presents the Exploratory Data Analysis (EDA) in detail. Section 4, presents the experimental results with the performance analysis and Section 5; discusses the conclusion.

II. DATASET DESCRIPTION

This dataset was collected from the website http://lib.stat.cmu.edu/datasets/bodyfat. It consists of 252 instances, 14 attributes (Independent Variables), and one dependent variable. Figure 1 shows the attribute description of the body fat dataset.

Independent Variables	Notation	Units
Density	Density	gm/cm3
Age	Age	years
Weight	Weight	lbs
Height	Height	inches
Neck Circumference	Neck	cm
Chest Circumference	Chest	cm
Abdomen Circumference	Abdomen	cm
Hip Circumference	Hip	cm
Thigh Circumference	Thigh	cm
Knee circumference	Knee	cm
Ankle Circumference	Ankle	cm
Biceps (extended) Circumference	Biceps	cm
Forearm Circumference	Forearm	cm
Wrist Circumference	Wrist	cm

Figure 1: Body Fat Dataset Description

2.1 Preprocessing

Normalization is a pre-processing technique that deals with the dataset in which there is a drastic difference in the attributes. The body fat dataset was processed with a linear transformation technique known as Min-Max Normalization. This technique ensures the equal priority among the attributes. The Min-Max Normalization is calculated using Equation 1. In Equation 1, D' represents normalized attribute value, Drepresents the original attribute value, Minrepresents minimum value, and Maxrepresents the maximum value of the attribute. By default, the value for new_max is one and the value for $new_min$is zero. Table 1, shows the sample dataset values before applying the normalization process. Table 2, shows the sample dataset values after applying the normalization process.

$$D' = \frac{D-Min}{Max-Min}(new\_max - new\_min) + new\_min \quad (1)$$

Table 1: Sample Dataset Values before Normalization

Density	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Bicep	Forearm	Wrist	Bodyfat
1.0708	23	154.25	67.75	36.2	93.1	85.2	94.5	59.0	37.3	21 .9	32.0	27.4	17.1	12.3
1.0853	22	173.25	72.25	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2	6.1
1.0414	22	154.00	66.25	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6	25.3
1.0751	26	184.75	72.25	37.4	101.8	86.4	101 .2	60.1	37.3	22.8	32.4	29.4	18.2	10.4
1.0340	24	184.25	71.25	34.4	97.3	100.0	101 .9	63.2	42.2	24.0	32.2	27.7	17.7	28.7
1.0502	24	210.25	74.75	39.0	104.5	94.4	107.8	66.0	42.0	25.6	35.7	30.6	18.8	20.9
1.0549	26	181.00	69.75	36.4	105.1	90.7	100.3	58.4	38.3	22.9	31.9	27.8	17.7	19.2
1.0704	25	176.00	72.50	37.8	99.6	88.5	97.1	60.0	39.4	23.2	30.5	29.0	18.8	12.4
1.0900	25	191.00	74.00	38.1	100.9	82.5	99.9	62.9	38.3	23.8	35.9	31.1	18.2	4.1

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

1.0722	23	198.25	73.50	42.1	99.6	88.6	104.1	63.1	41.7	25.0	35.6	30.0	19.2	11.7

III. EXPLORATORY DATA ANALYSIS (EDA)

The best way to visualize the distribution of the data is by using histograms, boxplots, and correlation matrix plot. In this paper, the EDA is applied to the normalized data. The histogram distribution of 15 independent variables are shown in Figure 2. In Figure 2, attributes such as body fat, density, and wrist data are normally distributed. Box-Plot is used to visualize and verify the quality of data. Figure 3 shows the box plot for 15 independent variables such as density, bodyfat, weight, height, and neck, etc. In the same figure, some of the data points value falls outside the whisker that is an abnormal value called an outlier. In this plot, density & body fat contains one upper outlier, and the wrist & chest contain two upper outliers, etc. Correlation Matrix Plot can show which variable having a high and low correlation with respect to another variable. The Correlation Plot is shown in Figure 4 in which Density, Abdomen, and Chest highly correlated with the body fat whereas the height, age and ankle are weakly correlated with the body fat. The [16] [17] Pearson Correlation Coefficient (PCC) used to quantify the linear relationship between two variables. The PCC measure is computed using Equation 2. In Equation (2), R represents the correlation coefficient, $\bar{x}$ represents the average value for all x values, $\bar{y}$ represents the average value for all y values and $x_i \& y_i$ represent the variables.

$$R = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i (x - \bar{x})^2} \sqrt{\Sigma_i (y - \bar{y})^2}} \quad (2)$$

Table 2: Dataset after Normalization

Density	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist	Bodyfat
0.665	0.017	0.146	0.793	0.254	0.243	0.201	0.152	0.294	0.267	0.189	0.356	0.460	0.232	0.259
0.793	0.000	0.224	0.886	0.368	0.251	0.173	0.219	0.287	0.267	0.291	0.282	0.568	0.429	0.128
0.407	0.000	0.145	0.762	0.144	0.290	0.235	0.226	0.309	0.366	0.331	0.198	0.302	0.143	0.533
0.703	0.068	0.271	0.886	0.313	0.395	0.216	0.258	0.322	0.267	0.250	0.376	0.604	0.429	0.219
0.342	0.034	0.269	0.865	0.164	0.316	0.389	0.270	0.399	0.571	0.331	0.366	0.482	0.339	0.604
0.485	0.034	0.375	0.938	0.393	0.443	0.318	0.364	0.469	0.559	0.439	0.540	0.691	0.536	0.440
0.526	0.068	0.255	0.834	0.264	0.453	0.271	0.244	0.279	0.329	0.257	0.351	0.489	0.339	0.404
0.662	0.051	0.235	0.891	0.333	0.357	0.243	0.193	0.319	0.398	0.277	0.282	0.576	0.536	0.261
0.834	0.051	0.296	0.922	0.348	0.380	0.166	0.238	0.392	0.329	0.318	0.550	0.727	0.429	0.086
0.678	0.017	0.326	0.912	0.547	0.357	0.244	0.305	0.397	0.540	0.399	0.535	0.647	0.607	0.246

IV. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

Regression analysis is one of the predictive modelingtechniques. The multiple regression analysis with n independent variables is computed using Equation 3.

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (3)$$

In equation (3), $\hat{Y}$represents the estimated value of the dependent variable, β_0 represents intercept, $\beta_1, \beta_2, \ldots \beta_n$represents the slope coefficient,$x, x_1 x_2 , \ldots x_n$ represents the independent variables, $n = 1,2, \ldots n$, and ε represents the random error. The difference between the actual value of the dependent variable and the value of the dependent variable predicted by the regression line is called error.

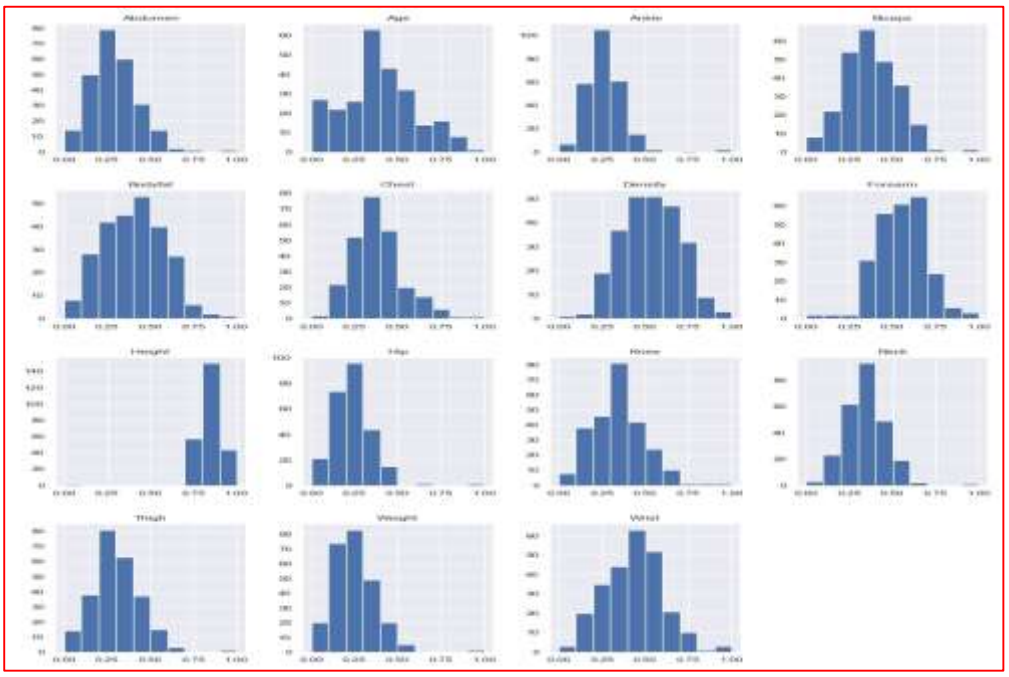

Figure 2: Histogram Distribution for Independent Variables

2.2 Regression Equation

Model-I: Multiple Regression Analysis (All Variables)

In this model, all the variables are included to predict body fat. The predicted regression line is shown in Equation 4.

$$\text{Bodyfat} = 0.884 - (0.970 * \text{Density}) + (0.018 * \text{Age}) + (0.067 * \text{Weight}) - (0.008 * \text{Neck}) + (0.036 * \text{Chest}) + (0.043 * \text{Abdomen}) - (0.011 * \text{Hip}) - (0.002 * \text{Thigh}) - (0.004 * \text{knee}) - (0.029 * \text{Ankle}) - (0.021 * \text{Biceps}) + (0.011 * \text{Forearm}) \quad (4)$$

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

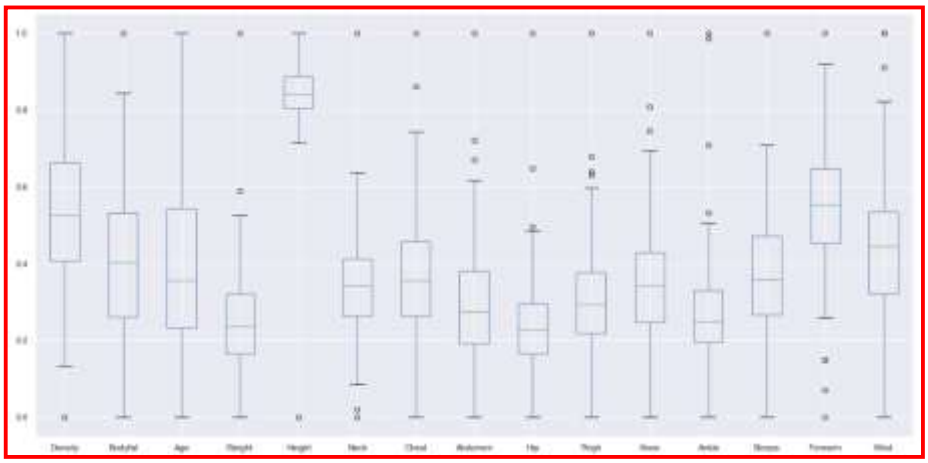

Figure 3: Box Plot for Independent Variables

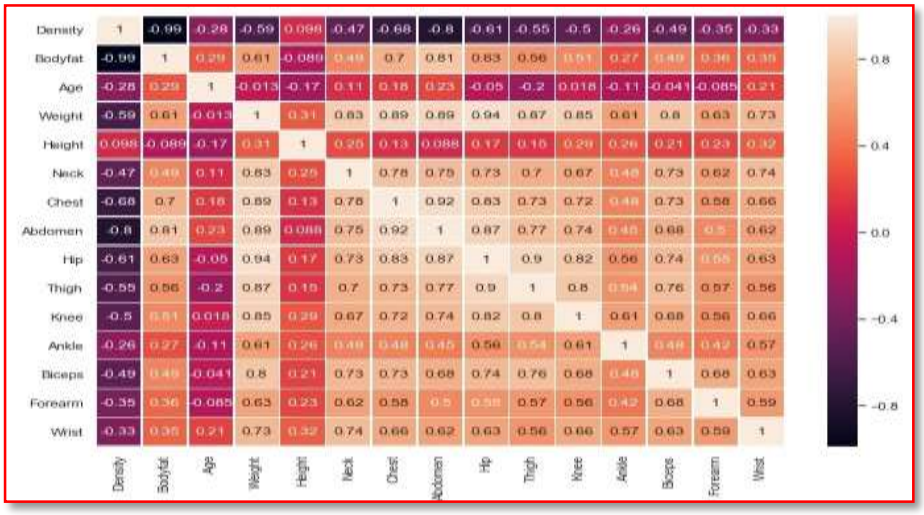

Figure 4: Correlation Plot for Independent Variables

Model-II: Multiple Regression Analysis (Significant Variables)

In this model, only significant variables such as density, thigh, knee, biceps, forearm, and wrist are included. The predicted regression line is shown in Equation 5.

$$\text{Bodyfat} = 0.922 - (1.015 * \text{Density}) + (0.006 * \text{Thigh}) + (0.022 * \text{Knee}) - (0.014 * \text{Biceps}) + (0.006 * \text{Forearm}) + (0.033 * \text{Wrist}) \quad (5)$$

Model-III: Support Vector Machine Regression Analysis

In this model, significant variables such as density, thigh, knee, biceps, forearm, and wrist are included. The predicted regression line is shown in Equation 6.

$$\text{Bodyfat} = 0.672 - (0.733 * \text{Density}) + (0.103 * \text{Thigh}) + (0.133 * \text{Knee}) - (0.034 * \text{Biceps}) + (0.093 * \text{Forearm}) + (0.009 * \text{Wrist}) \quad (6)$$

2.3 Regression Assumption Comparison

The coefficient of determination (R^2) is a goodness of fit measure of the regression that explains how well the model fits the data. The value lies between 0 to 1. The interpretations of model fit depends on the context of

analysis. The regression model comparisons results are shown in Table 3. Model 1 & 2 have the same coefficient of determination value than the Model 3. In linear regression, the model should not suffer from Auto Correlation, Multicollinearity, and Hetroscedacity problem. Autocorrelation occurs when the errors are dependent on each other. The Durbin-Watson Statistic (DW) test was used to test the autocorrelation in regression analysis. The normal value for Durbin-Watson Statistic is 1.5 to 2.5. Multicollinearity occurs when some or all of the independent variables are highly related to one another. Variance Inflation Factor (VIF) used to test multicollinearity in regression analysis. The Multicollinearity exists in the model if the VIF value exceeds more than 10. The error terms must have constant variance is referred to as homoskedasticity and non-constant variance is referred to as heteroscedasticity. Breusch-Pagan (BP) test was used to test heteroscedasticity in a linear regression analysis. The null hypothesis states that the variances are equal. The alternative hypothesis states that variances are unequal. If the p-value is less than significance level of 0.05 then it is rejected as the null hypothesis. The model comparison result chart is shown in Figure 5.

Table 3: Model Comparison Result

	Model 1	Model 2	Model 3
R^2	0.946	0.946	0.804
DW Test	1.972	1.979	2.046
VIF Test	19.039	18.503	9.229
BP Test	23.9299	8.890	15.634
p-value	0.021	0.114	0.008

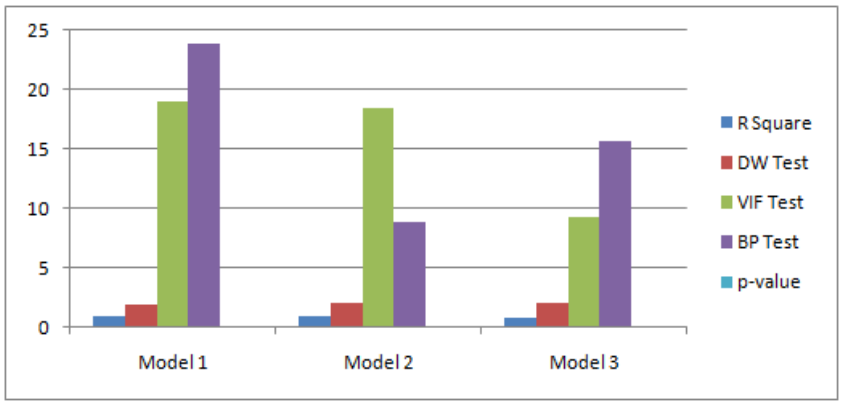

Figure 5: Model Comparison Result chart

4.3 Regression Model Validation

The model is validated based on Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). The MAE, RMSE, and MSE values for the regression model are shown in Table 4. Model 1 & 2 has less error value than Model 3. The model validation comparison chart shown in Figure 6.

Table 4: Regression Model Validation

	Model 1	Model 2	Model 3
MAE	0.0097	0.0097	0.047
RMSE	0.044	0.044	0.062
MSE	0.002	0.002	0.004

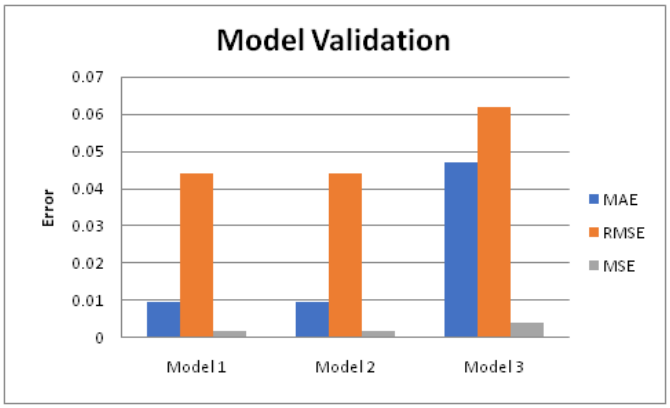

Figure 6: Model Validation Comparison Chart

V. CONCLUSION

In this paper, regression models such as multiple regression with all variables, multiple regression with significant variables and support vector regression are analysed to predict the body fat. The regression equations are formulated to predict the body fat. The goodness of measure the coefficient of determination (R^2) values are calculated and compared. The R^2 value for Model 1& 2 is 0.946. These two models are better than the support vector regression model. The linear regression model assumptions such as Hetroscedacity, Auto Correlation, and Multicollinearity are checked. Finally, the model is validated based on Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). Model 1 & 2 have less error than Model 3. In future, this model can be implemented in real-time medical applications.

REFERENCES

1. Shiva Shanth Reddy Ainala.: Study on body fat density prediction based on anthropometric variables. International Journal of Data Mining & Knowledge Management Process, vol 5(3), pp 1--8(2015).
2. J.P. Verma, J.P. Bhukar, Prasenjeet Biswas.: Models in Estimating Fat Percentage in Active Male. European Journal of Physical Education and Sport, vol.13(3), pp 103--107(2016).
3. St-Onge MP.: Relationship between body composition changes and changes in physical function and metabolic risk factors in aging. Current Opinion Clinical Nutrition Metabolic Care, vol. 8, pp 523--528(2005).
4. Martarelli D, Martarelli B, Pompei P.: Body composition obtained from the body mass index: an Italian study. European Journal of Nutrition, vol. 47, pp 409--416(2008).
5. Gómez-Ambrosi J, Silva C, Catalán V, Rodríguez A, Galofré JC, Escalada J.: Clinical usefulness of a new equation for estimating body fat. Diabetes Care. vol. 35, pp 383--388(2012).

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

6. Jackson AS, Stanforth PR, Gagnon J, Rankinen T, Leon AS, Rao DC.: The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. International Journal of Obesity, vol 26, pp 789--796(2002).
7. Burton BT, Foster WR, Hirsch J, Van Itallie TB: Health implications of obesity: an NIH Consensus Development Conference. International Journal of Obesity. vol. 9, pp 155--170(1985).
8. Chathuranga Ranasinghe, Prasanna Gamage, Prasad Katulanda, Nalinda Andraweera, Sithira Thilakarathne, and Praveen Tharanga.: Relationship between Body mass index (BMI) and body fat percentage, estimated by bioelectrical impedance, in a group of Sri Lankan adults: a cross sectional study. BMC Public Health, vol 13(1), pp 797--805(2013).
9. Ifeoma, F. Odo, Lawrence, U. S. Ezeanyika and Nene, Uchendu.: The Relationship among Body Composition and Body Mass index in a Population of Adolescents in Enugu State, Nigeria. International journal of current Microbiology and applied science, vol. 4(1), pp 884--897(2015).
10. Pilar Fuster-Parra, Miquel Bennasar-Veny, Pedro Tauler, Aina Yañez, AngelA. López-González, Antoni Aguiló.: A Comparison between Multiple Regression Models and CUN-BAE Equation to Predict Body Fat in Adults, Plos One (2015).
11. Su-Hsin Chang, Tracey S. Beason ,Jean M. Hunleth and Graham A. Colditz,: A systematic review of body fat distribution and mortality in older people. Maturitas The European Menopause Journal, vol. 72(3), pp 175--191(2012).
12. Manish S. Bharadwaj, Daniel J. Tyrrell, Iris Leng, Jamehl L. Demons, Mary F. Lyles, J. Jeffrey Carr, Barbara J. Nicklas, and Anthony J. A. Molina.: Relationships between mitochondrial content and bioenergetics with obesity, body composition and fat distribution in healthy older adults. BMC Obesity, vol. 2, pp 40--51(2015).
13. Bandana Sen, Kaushik Bose, Saijuddin Shaikh, and Dilip Mahalanabis.: Prediction Equations for Body-fat Percentage in Indian Infants and Young Children Using Skinfold Thickness and Mid-arm Circumference. Journal of health, population, and nutrition, vol.28, pp221--229(2010).
14. Manuel Ramirez-Zea, Benjamin Torun, Reynaldo Martorell, and Aryeh D Stein.: Anthropometric predictors of body fat as measured by hydrostatic weighing in Guatemalan adults. American journal of clinical nutrition, vol.83,pp795--802(2006).
15. Pongchaiyakul, Chatlert and Kosulwat, Vongsvat and Charoenkiatkul, Somsri and Thepsuthammarat, Kaewjai and Nguyen, Tuan and Rajatanavin, Rajata.: Prediction of Percentage Body Fat in Rural Thai Population Using Simple Anthropometric Measurements. Obesity research, vol.33, pp729--738(2005).
16. S. Velliangiri, P. Karthikeyan, I. T. Joseph and S. A. P. Kumar, "Investigation of Deep Learning Schemes in Medical Application," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, 2019, pp. 87-92.
17. S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar Joseph, “A Review of Dimensionality Reduction Techniques for Efficient Computation,” Procedia Comput. Sci., vol. 165, pp. 104–111, 2019.