Intelligent Fat Predictor: Leveraging Linear Regression and K Nearest Neighbors in Obesity diseases.

Mona Mohamed

Higher Technological Institute, 10th of Ramadan City 44629, Egypt

Email: mona.fouad@hti.edu.eg

Abstract

One of the major lifestyle disorders brought on by unwholesome daily routines and inherited ailments is obesity and overweight. And this illness is a risk factor for a wide range of chronic illnesses, such as cancer, diabetes, metabolic syndrome, and cardiovascular conditions. Additionally, according to the World Health Organization (WHO), 30% of deaths worldwide will be caused by lifestyle illnesses by 2030. These deaths can be prevented by appropriately identifying and treating risk factors that relate to these diseases as well as by implementing behavioral engagement policies. Thence, the study is leveraging machine learning (ML) techniques for analyzing data and discovering new patterns for predicting body fat. The problem of predicting fat classifies as a regression, hence, we are deploying two regression techniques to deal with the regression dataset. These techniques are used linear regression (LR) and k nearest neighbors (KNN) which fall under umbral of ML. The two techniques are applied on real datasets. The dataset has 252 records. The results showed the LR has the highest score than the KNN model.

Keywords: Machine Learning; Linear Regression; K Nearest Neighbors; Body Fat; Prediction; Regression Problem, obesity.

1. Introduction and problem setting

Obesity is a substantial wellness issue, as scholars in [1] indicated that it is the fifth largest cause of mortality worldwide. In a similar vein [2], where the nationwide epidemic of obesity that is presently impacting the pediatric population has emerged as a significant global health issue. World Health Organization (WHO) depicted obesity in [3] as a buildup of fat that is abnormal or excessive and might be harmful to health, due to calories ingested and calories burned are not balanced out. expended. Evidence that this is true Ref [4] Obesity is a metabolic disorder marked by uncontrollable weight gain brought on by an excessive proportion of fat, a high-calorie intake, and a lack of energy expenditure. Whilst the most up-to-date assessments of global obesity rates in [5] a minimum of 30% of men and 35% of women are obese in a number of regions throughout the globe.

According to the study's findings [6],[7] obesity raises the risk of contracting a variety of chronic disorders and illnesses, such as asthma, cancer, diabetes, high blood cholesterol, and cardiovascular disease. A study conducted by de Siqueira et al. [8] explored and appraised any possible associations between obesity and COVID-19. The study's final verdict claimed that obesity was a detrimental factor for COVID-19 based on the research that was done. A higher Body Mass Index (BMI) in particular had a poorer outcome, which was established by taking into account higher mortality rates, higher hospitalization rates, and worse prognosis and recovery outcomes.

Given the seriousness of obesity and other problems, as announced in previous survey studies, it is essential to be able to predict body fat percentage when making a diagnosis of obesity. These motives strongly influenced the decision to carry out this study. A synopsis of the challenges that the study was concentrating on can be seen in Figure 1.

Scholars as Ref [9] attempted to solve mentioned problems in Figure 1 through deploying computational intelligent techniques like machine learning (ML) techniques as subset of artificial intelligence (AI) for predicting body fat. In

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

order to ML can increase comprehension of obesity and the potential for prediction with never-before-seen precision. Due to ML's potential for characterizing, adapting, learning, predicting, and analyzing data, a greater knowledge of obesity and the ability to predict future events [10].

For instance, Artificial Neural Network (ANN) model as one of the most popular technique of ML which applied in [11] which extraordinarily precise estimation of body fat percentages. In similar vein Shao et al. [12] predicted percentage of body fat through volunteering support vector machine (SVM).

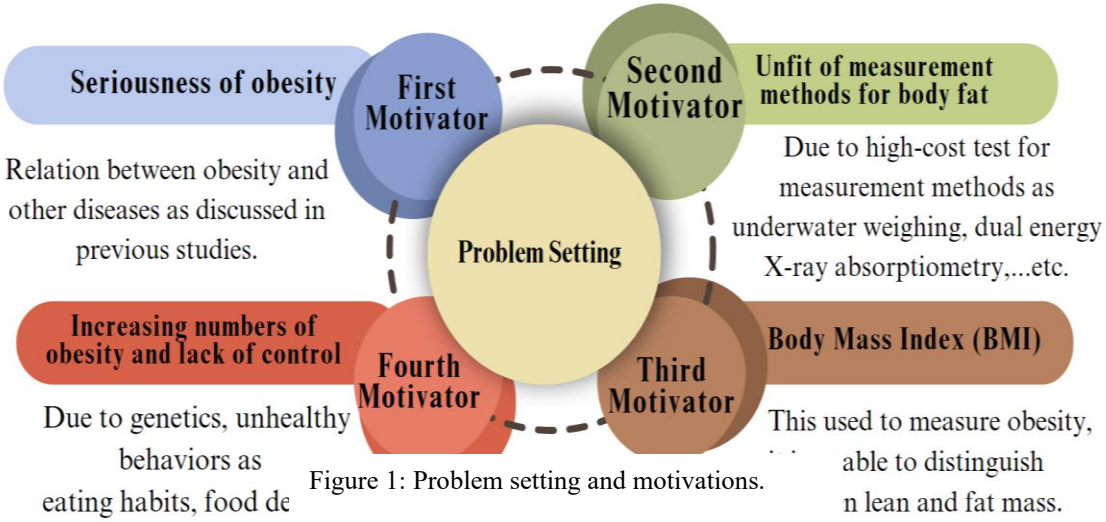

Figure 1: Problem setting and motivations.

Hence, this study invests the benefits of applying ML techniques to predict fat body. Herein, we volunteered two ML techniques for body fat prediction. These techniques encompassed linear regression (LR), and K nearest neighbors (KNN).

The study consists of 4 main sections each one responsible for certain information and process as section two which provides findings for systematic bibliometrics analysis, while the utilized datasets and techniques which applied in our methodology are clarified in section 3. So, the results of application of ML techniques on dataset are obtained in section 4. Finally, our conclusion is represented in section 5.

2. Systematic Analysis for study

In this section we track steps in [13] for conducting Bibliometrics analysis process. Hence, we conduct surveys on earlier studies and the result of this analysis can be shown in mapping networks. Firstly, we determine the keywords which are utilized in queries based on web of science (WoS) database. These keywords represented as ("Fat prediction" AND "Machine Learning" And "Obesity" during 2015 until 2023. Secondly, we are applying VOS viewer for generating mapping network for determined queries. Whereas Figure 2 is finding of analyzing co-occurrences for All Keywords. Relying on this Figure there are 11 items which clustering into 3 clusters with 101 Total Length Strength (TLS). Cluster one includes 6 items while cluster 2 has 4 items and cluster 3 has 1 item.

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

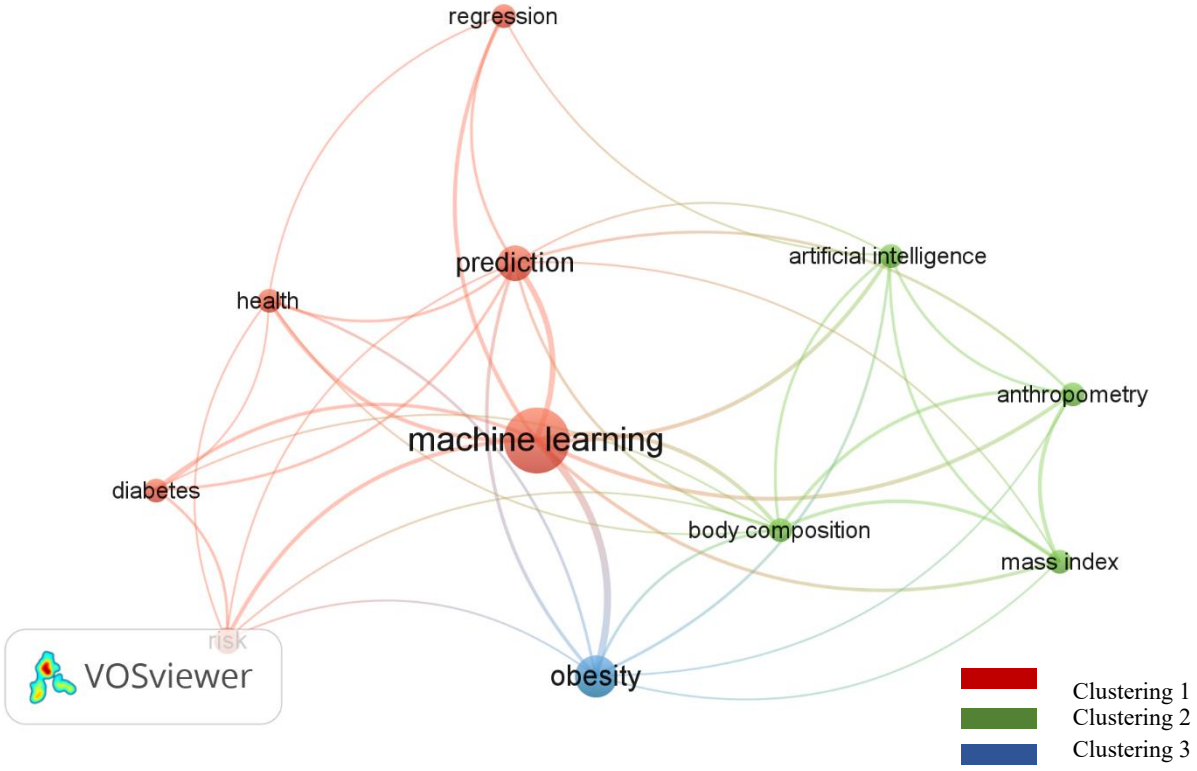

Figure 2: Co-Citation analysis network of machine learning in fat prediction / Obesity

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

3. Material and Methods

Herein, we describe the dataset which volunteers in our proposed methodology. also, we are elucidating the techniques of ML which contribute to constructing our study.

3.1 Dataset Description

The dataset of body fat is collected from Kaggle. Table 1 shows the sample dataset. The dataset has 252 records and 14 features. The dataset's features are shown in the first row in Table 1. The distribution of features in the dataset are clarified in Figures 3-6.

Table 1: The sample of body fat dataset

	Density	BodyFat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps
BFP0	1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59	37.3	21.9	32
BFP1	1.0853	6.1	22	173.25	72.25	38.5	93.6	83	98.7	58.7	37.3	23.4	30.5
BFP2	1.0414	25.3	22	154	66.25	34	95.8	87.9	99.2	59.6	38.9	24	28.8
BFP3	1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4
BFP4	1.034	28.7	24	184.25	71.25	34.4	97.3	100	101.9	63.2	42.2	24	32.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
BFP247	1.0736	11	70	134.25	67	34.9	89.2	83.6	88.8	49.6	34.8	21.5	25.6
BFP248	1.0236	33.6	72	201	69.75	40.9	108.5	105	104.5	59.6	40.8	23.2	35.2
BFP249	1.0328	29.3	72	186.75	66	38.9	111.1	111.5	101.7	60.3	37.3	21.5	31.3
BFP250	1.0399	26	72	190.75	70.5	38.9	108.3	101.3	97.8	56	41.6	22.7	30.5

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

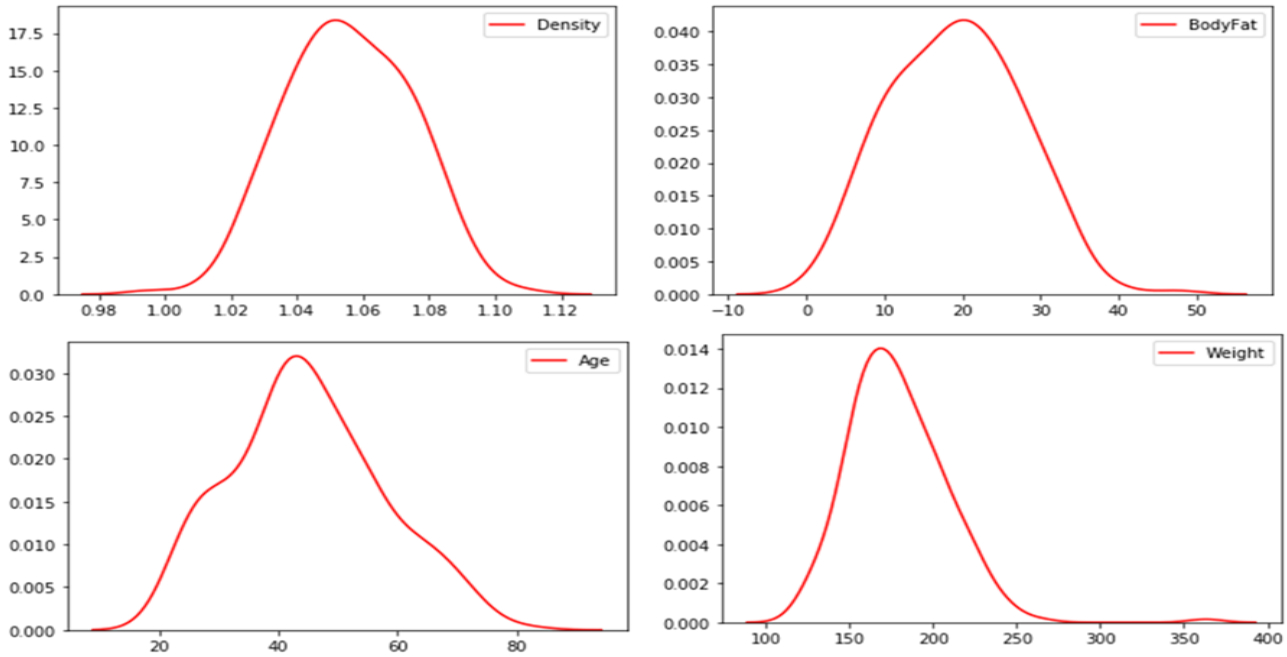

Figure 3: The density, age, body fat, and weight features.

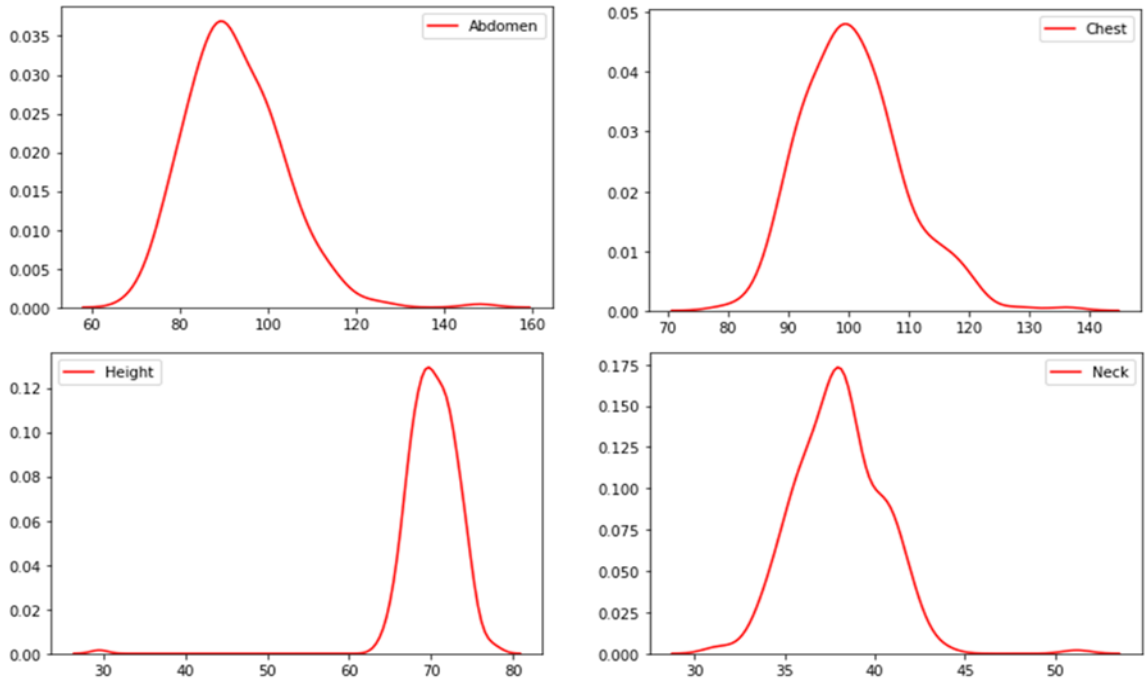

Figure 4: The abdomen, chest, Neck, and height features.

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

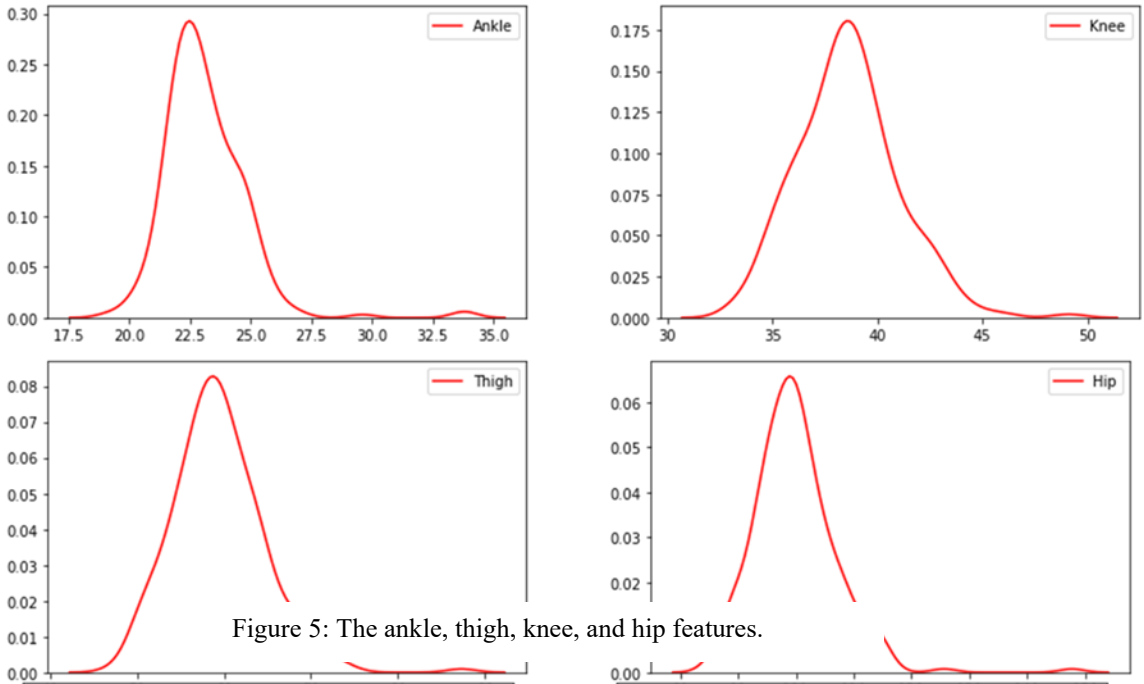

Figure 5: The ankle, thigh, knee, and hip features.

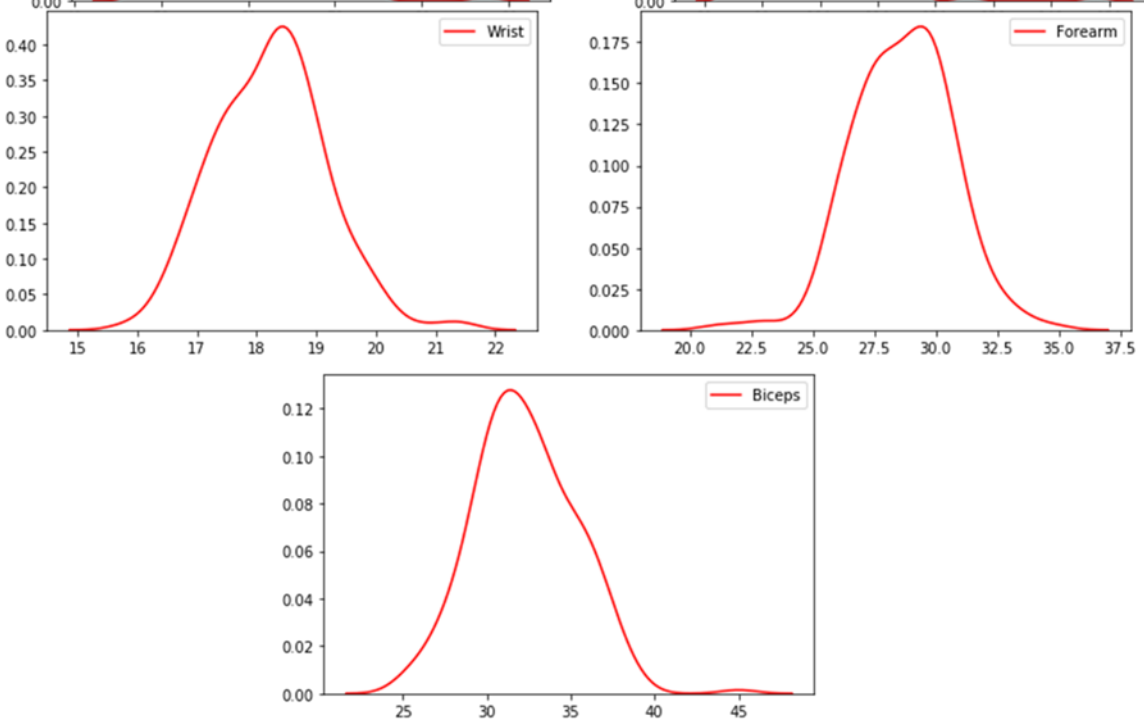

Figure 6: The wrist, biceps, and forearm features.

3.2 Prediction Techniques

This subsection encompasses utilized ML techniques which contributed to conducting our study.

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

3.2.1 Linear Regression

Multiple linear regression (LR) is a parametric technique used to determine the existence of linear relationships among predictor and predicted parameters[14],[15],[16].

$$E_r = C_0 + \sum_{i=1}^{m} C_i y_{ir} \quad (1)$$

Where E is an estimated variable and C is a coefficient of predictors.

3.2.2 K -Nearest Neighbors (KNN)

Without specifying a parametric link among the predictor and forecasted parameters beforehand, K-NN regression allows for real-time forecasting of the quantity of the predicted value using information gleaned from the data being collected [17],[18],[19].

This technique is based on using the Euclidean distance function (S_{rt}) to determine the proximity (neighborhood) between the number of indicators for each historical assessment.

Y_t = y_1t,y_2t,y_3t,…….y_mt and the number of indicators for every current assessment Y_r = y_1r,y_2r,y_3r,…….y_mr

$$S_{rt} = \sqrt{\sum_{i=1}^{m} w_i (y_{ir} - y_{it})^2} \quad (2)$$

The forecasted value can be computed by using function of probabilistic as:

$$P_t = \sum_{j=1}^{K} q(S_{rj}) \times T_j \quad (3)$$

The kernel function can be computed as:

$$q(S_{rj}) = \frac{\frac{1}{S_{rj}}}{\sum_{j=1}^{K} \frac{1}{S_{rj}}} \quad (4)$$

4. Results and Discussion

This section presented the results of applying the two models in the body fat dataset. Table 2 clarified the descriptive statistics on the dataset. The dataset has 252 cases and 15 features. Feature 4 has the highest average, standard deviation, and maximum value. Feature 2 has the minimum value

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

This study-built preprocessing on the dataset. In weight features, this paper selected weights greater than 250. In the height feature, this study selected a height less than 30. Figures 7 and 8 show the weight and height after preprocessing dataset respectively.

Table 2: The statistics analysis of dataset.

	Amount	Average	Standard deviation	minimum	25%	50%	75%	Maximum
BFC1	252	1.055574	0.019031	0.995	1.0414	1.0549	1.0704	1.1089
BFC2	252	19.15079	8.36874	**0**	12.475	19.2	25.3	47.5
BFC3	252	44.88492	12.60204	22	35.75	43	54	81
BFC4	252	**178.9244**	**29.38916**	118.5	159	176.5	197	**363.15**
BFC5	252	70.14881	3.662856	29.5	68.25	70	72.25	77.75
BFC6	252	37.99206	2.430913	31.1	36.4	38	39.425	51.2
BFC7	252	100.8242	8.430476	79.3	94.35	99.65	105.375	136.2
BFC8	252	92.55595	10.78308	69.4	84.575	90.95	99.325	148.1
BFC9	252	99.90476	7.164058	85	95.5	99.3	103.525	147.7
BFC10	252	59.40595	5.249952	47.2	56	59	62.35	87.3
BFC11	252	38.59048	2.411805	33	36.975	38.5	39.925	49.1
BFC12	252	23.10238	1.694893	19.1	22	22.8	24	33.9
BFC13	252	32.27341	3.021274	24.8	30.2	32.05	34.325	45
BFC14	252	28.66389	2.020691	21	27.3	28.7	30	34.9
BFC15	252	18.22976	0.933585	15.8	17.6	18.3	18.8	21.4

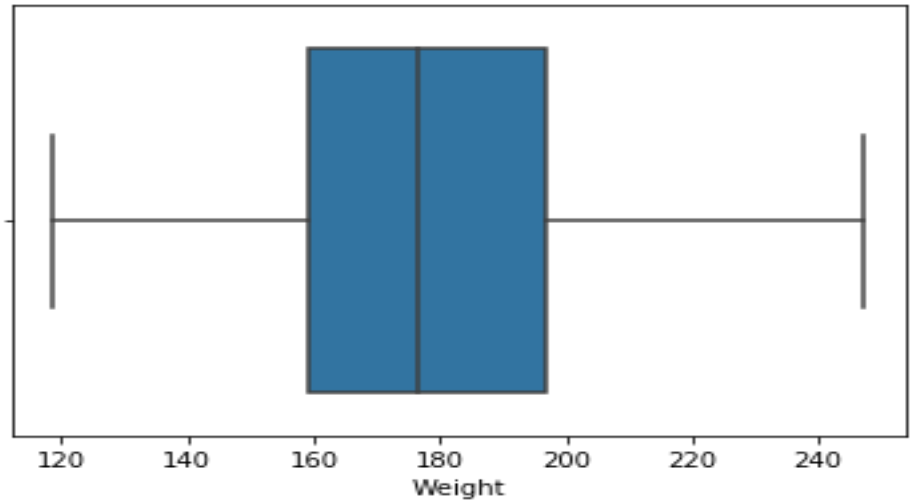

Figure 7: The weight feature after preprocessing dataset.

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

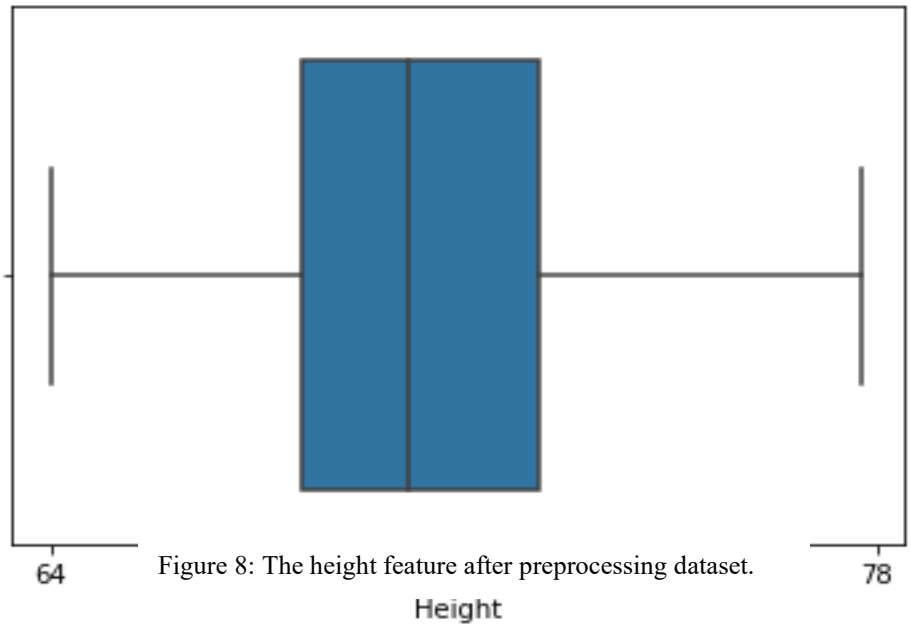

Figure 8: The height feature after preprocessing dataset.

The score of two techniques into dataset is illustrated via Figure 9. Relying on this Figure, we observed that LR has the highest score 0.91 while KNN has the lowest score 0.545. Hence, based on the findings of score in Figure 8 we acknowledge LR's ability comparing to KNN to predict body fat based on this dataset.

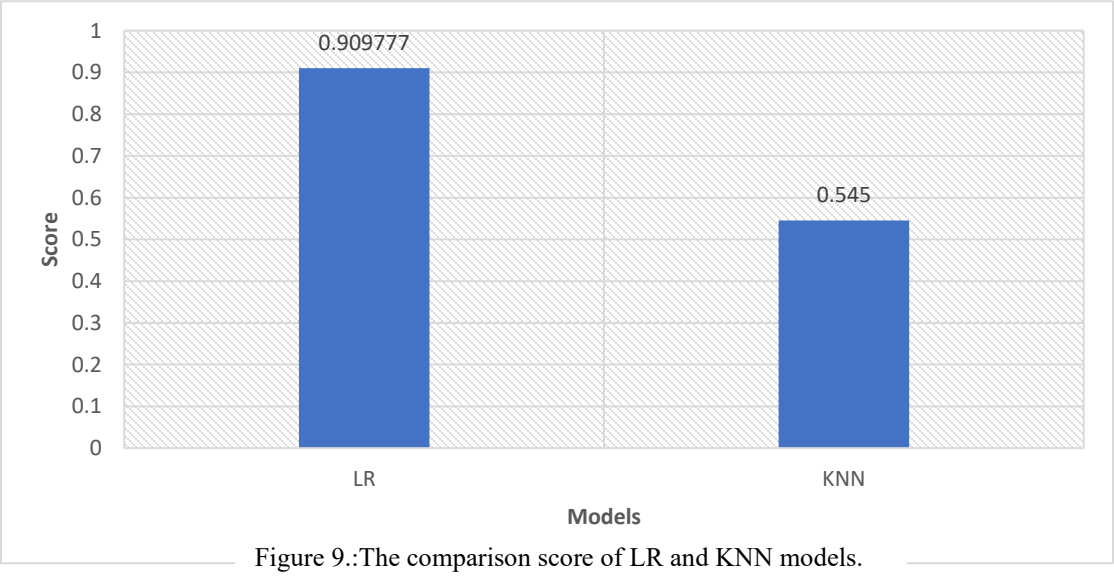

Figure 9.:The comparison score of LR and KNN models.

5. Conclusion

Relying on conducted surveys to reveal studies about obesity, there are multiple manners in which obesity is detrimental to human health and exposure to risks. Thus, early detection of obesity is important. Consequently, Studies have thus worked hard to pinpoint the early causes.

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

This study attempted to realize the shortcomings of existing measurement methods for estimating and measuring fats in the body. Through the utilization of potent techniques like random forest (RF), ANN, KNN…etc. This study explores the capability of computational intelligence approaches for performing exploratory analysis of data to uncover patterns or behaviors on information.

Herein, utilized ML techniques in this study are promising approach to early predictions of obesity and the risk of overweight since it can provide rapid. Thus, LR, and KNN have volunteered to serve the objective of our study through applying these techniques on real dataset to predict body fat. Guided by the findings of the two techniques' experiments and Figure 8, we concluded that LR outperforms KNN through applying each one on the described dataset.

Reference

[1] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila, and A. Shapi'i, "A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity," *Comput. Biol. Med.*, vol. 136, no. April, p. 104754, 2021, doi: 10.1016/j.compbiomed.2021.104754.

[2] P. Costa-Urrutia *et al.*, "Obesity measured as percent body fat, relationship with body mass index, and percentile curves for Mexican pediatric population," *PLoS One*, vol. 14, no. 2, pp. 1–13, 2019, doi: 10.1371/journal.pone.0212792.

[3] S. Camacho and A. Ruppel, "Is the calorie concept a real solution to the obesity epidemic?," *Glob. Health Action*, vol. 10, no. 1, 2017, doi: 10.1080/16549716.2017.1289650.

[4] G. Hernández, "Prevalencia de sobrepeso y obesidad, y factores de riesgo, en niños de 7-12 años, en una escuela pública de Cartagena," *Colombia–2010*, 2011.

[5] L. A. Fowler, J. R. Fernández, S. E. Deemer, and B. A. Gower, "Genetic risk score prediction of leg fat and insulin sensitivity differs by race/ethnicity in early pubertal children," *Pediatr. Obes.*, vol. 16, no. 12, pp. 2–9, 2021, doi: 10.1111/ijpo.12828.

[6] L. Hu *et al.*, "Prevalence of overweight, obesity, abdominal obesity and obesity-related risk factors in southern China," *PLoS One*, vol. 12, no. 9, pp. 1–14, 2017, doi: 10.1371/journal.pone.0183934.

[7] N. Young, I. K. Atan, R. G. Rojas, and H. P. Dietz, "Obesity: how much does it matter for female pelvic organ prolapse?," *Int. Urogynecol. J.*, vol. 29, no. 8, pp. 1129–1134, 2018, doi: 10.1007/s00192-017-3455-8.

[8] J. V. V. de Siqueira, L. G. Almeida, B. O. Zica, I. B. Brum, A. Barceló, and A. G. de Siqueira Galil, "Impact of obesity on hospitalizations and mortality, due to COVID-19: A systematic review," *Obes. Res. Clin. Pract.*, vol. 14, no. 5, pp. 398–403, 2020.

[9] K. W. DeGregory *et al.*, "A review of machine learning in obesity," *Obes. Rev.*, vol. 19, no. 5, pp. 668–685, 2018, doi: 10.1111/obr.12667.

[10] R. L. Figueroa and C. A. Flores, "Extracting information from electronic medical records to identify the obesity status of a patient based on comorbidities and bodyweight measures," *J. Med. Syst.*, vol. 40, pp. 1–9, 2016.

[11] A. Kupusinac, E. Stokić, and R. Doroslovački, "Predicting body fat percentage based on gender, age and BMI by using artificial neural networks," *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 610–619, 2014.

[12] Y. E. Shao, "Body fat percentage prediction using intelligent hybrid approaches," *Sci. World J.*, vol. 2014, 2014.

[13] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *J. Bus. Res.*, vol. 133, pp. 285–296, 2021.

[14] S. Kohli, G. T. Godwin, and S. Urolagin, "Sales prediction using linear and KNN regression," in *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019*, 2021, pp. 321–329.

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022

[15] T. Adithiyaa, D. Chandramohan, and T. Sathish, "Optimal prediction of process parameters by GWO-KNN in stirring-squeeze casting of AA2219 reinforced metal matrix composites," *Mater. Today Proc.*, vol. 21, pp. 1000–1007, 2020.

[16] A. Khazaee Poul, M. Shourian, and H. Ebrahimi, "A comparative study of MLR, KNN, ANN and ANFIS models with wavelet transform in monthly stream flow prediction," *Water Resour. Manag.*, vol. 33, pp. 2907–2923, 2019.

[17] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augment. Hum. Res.*, vol. 5, pp. 1–16, 2020.

[18] H. I. Dino and M. B. Abdulrazzaq, "Facial expression classification based on SVM, KNN and MLP classifiers," in *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 2019, pp. 70–75.

[19] S. Zheng and C. Ding, "A group lasso based sparse KNN classifier," *Pattern Recognit. Lett.*, vol. 131, pp. 227–233, 2020.

Doi: https://doi.org/10.54216/IJAACI.030101
Received: March 24, 2022 Revised: August 11, 2022 Accepted: January 05, 2022