CITY UNIVERSITY OF LONDON EST 1894

Predicting Obesity Level Using Random Forest and Logistic Regression

Syed Mahbubul Huq, MSc Data Science

Description and Motivation

Obesity is one of the main and growing problems. According to WHO Worldwide obesity has nearly tripled since 1975 and in 2016, more than 1.9 billion adults, 18 years and older, were overweight. Of these, over 650 million were obese [2]. This causes several health problems, which may lead to severity for human beings if not solved and predicted on time.

Most of teenagers have a deep connection with obesity and face its rising problems. If there had been a proper system that would have warned them or predicted they would be of great use for all ages of people. With this motivation in mind, this paper has done extensive analysis in applying two machine learning models, Random Forest and Logistic Regression in predicting obesity levels. Several pieces of research are conducted on this problem, De-La-Hoz-Correa being one who analyzed 3 machine learning models, Decision Tree (DT), Naïve Bayes (NB) and Logistic Regression (LR in predicting classifying the level of obesity on a similiar dataset which we are using [3].

Analysis of the Data

-The dataset contained, 2111 rows and 17 columns.

-Data gathered were generated were based on the people of Mexico, Peru and Colombia, 23% of the data was directly obtained from users via a web platform, while 77% of the data were generated synthetically using the Weka tool and the SMOTE filter [1].

-The dataset mainly had categorical data with many uneven float values in different rows, here and there. All of the categorical columns are thus rounded to their nearest integer value to avoid complexity.

-Label Encoding was done to make the dataset categorical Columns, Age, Weight and Height are also fixed so that they don't go to an infinite level after the decimal.

-All of the column headings are lastly modified to make in easier to look at, read and understand.

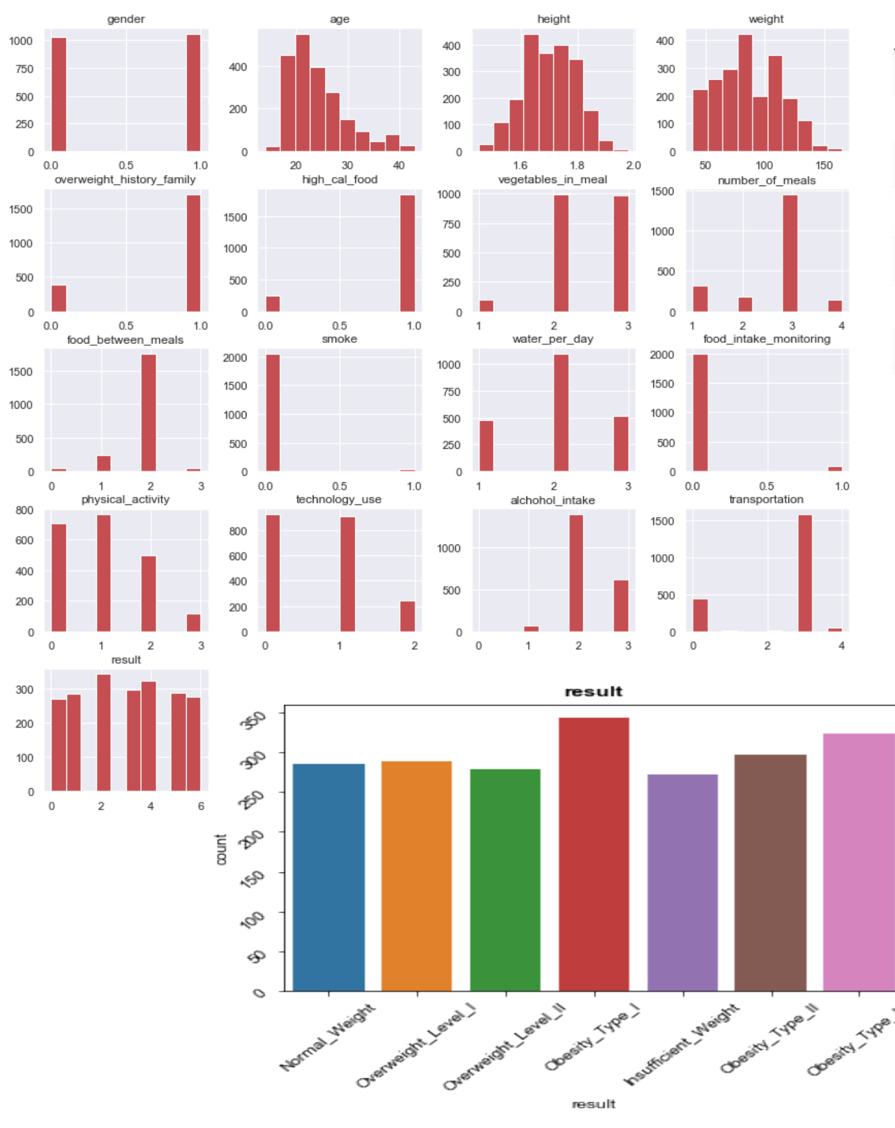
-Missing values and outliers are detected and removed from our dataset

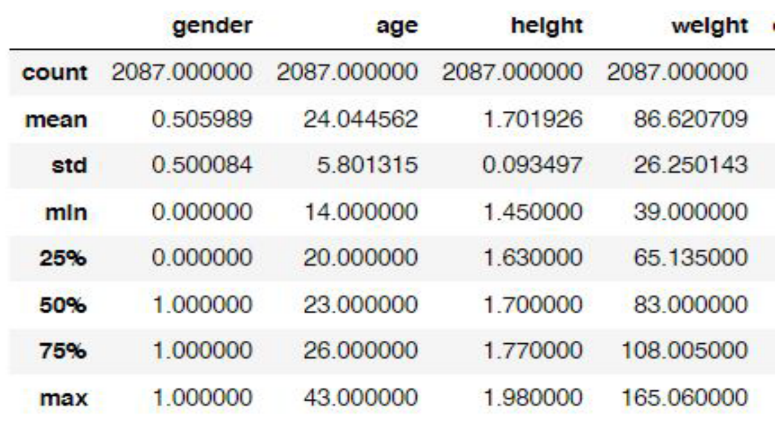
-Doing some basic analysis of data and visual representation, seemed to look like all the targeted results to predict are balanced with Obesity Type I dominating by a slight margin.

-A few of the columns are not balanced which would be fixed in the later stage

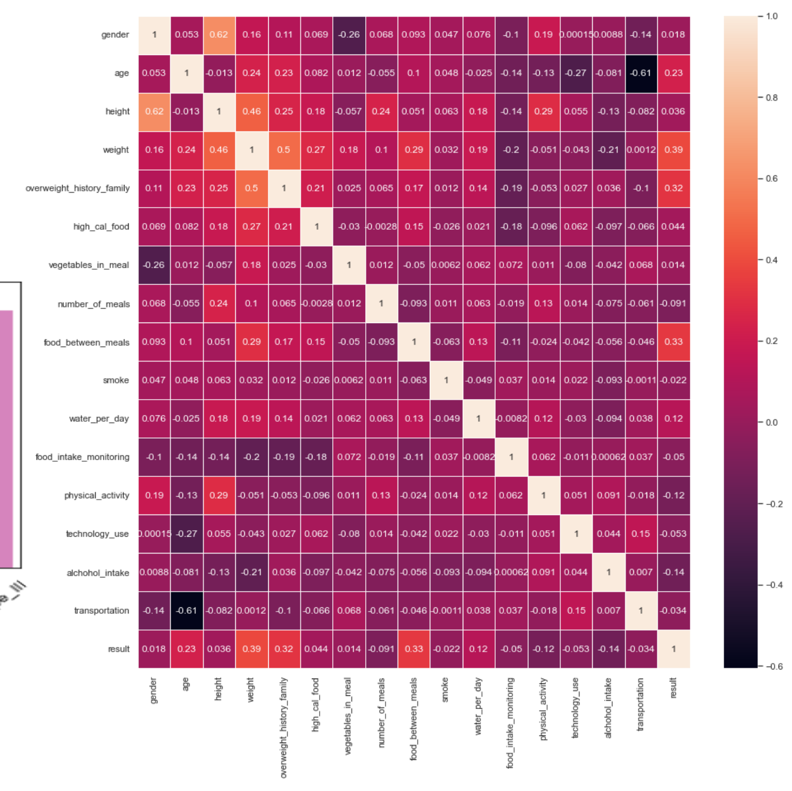
-A correlation heatmap is produced which shows the relationship of a column to other. It is observed very small columns, in height and weight are positively a bit correlated, this information would be vital to do feature reduction later on

-Looking at the statistical table about our data, it can be said most of the audience were young adults with a mean age of just 24 years with an average weight of 86 kg.





	gender	age	height	weight
count	2087.000000	2087.000000	2087.000000	2087.000000
mean	0.505989	24.044562	1.701926	86.620709
std	0.500084	5.801315	0.093497	26.250143
min	0.000000	14.000000	1.450000	39.000000
25%	0.000000	20.000000	1.630000	65.135000
50%	1.000000	23.000000	1.700000	83.000000
75%	1.000000	26.000000	1.770000	108.005000
max	1.000000	43.000000	1.980000	165.060000



Random Forest

In this algorithm, a combination of Decision Trees are made. As a result of that, it outperforms the decision tree in many instances. Though the computational cost increases, as a whole the accuracy increases. A paper [4] with a different dataset on predicting obesity had the highest accuracy rate in predicting obesity using RF, although the number of attributes was very high compared to ours. Besides, in another paper [5] RF performed really well after SVM and KNN with a completely different dataset. In a similar paper [6] that deals with detecting diabetes had the highest accuracy in detecting while using RF which is a bit similar to our project. These made me choose RF as in other papers this kind of algorithm showed good accuracy which might be the case in our project too.

Pros

-Accuracy increases

-Bias decreases and also avoids overfitting

-Works well with non-linear data[7].

-Follows a rule-based approach, so data normalization is not required.

Cons

-Increases computational power

-Requires more time to compute

-Sometimes, failing to determine the significance of each variable

Evaluation Methodology

-Mainly two datasets were used to compare the model performance. The first one is the main and original dataset which was label encoded, cleaned and preprocessed having 17 columns. The other dataset produced had feature reduction. Some of the imbalanced features were removed. Besides, highly correlated features were also reduced by deleting them. This includes the 'smoke', 'food_intake_monitoring' , 'high_cal_food' and 'overweight_history_family' columns. Both datasets were run on both of the models to see the result and outcome.

-Training and testing of the datasets were done with a ratio of 80% training and 20% testing data in all of the cases for both the models and datasets.

-For Random Forest 5-Fold Cross Validation was carried out to make the model more reliable and robust

-Hyperparameter Tuning was also carried out for Random Forest using the 'OptimizeHyperparameter' feature of MATLAB.

-A confusion matrix was made and also F-1 score, accuracy, and precision were calculated for all the models and variations to conclude our findings.

-MATLAB Classification Learner App was used to generate some useful insights from our analysis.

Analysis and Critical Evaluation of Results

-Although our second model, Logistic Regression did not run well and did not give us a good F1 score of 0.55 but accuracy was decent enough, but our first model Random Forest performed well and gave us an accuracy of 95.54% with F1 score of 0.95

-In the case of Random Forest, it was run as default with no tuning of parameters and also ran later on tuning it on two different datasets, the modified one and the normal one. Surprisingly, the model did not perform too well when modified datasets were used to do the modeling. An accuracy of 81% and F1 score of 0.81 was achieved while doing the modeling. This tells us much about Random Forest and its characteristics and behavior on our data. Reduction of feature and data normalization cannot be always a good option, which we can see in our case. Though in the case of

-Random Forest, it works as depth, so data normalization or standardization is not too helpful. Feature removal is also not too helpful and we can see evidence of that.

Random Forest when ran in default without doing any parameter tuning gives as an accuracy of 91%, after doing the hyperparameter tuning the accuracy increased and the efficiency of the result was achieved. It is an important feature for any machine learning model to do optimization as hyperparameter handle all the complexities of model and affects variance-base tradeoff [10]

-In case of logistic regression, a poorer F1 score but a relatively higher accuracy score of 81% was achieved. The F1 score was less compared to the accuracy because of the model being overfitted. New parameters were not balanced and a great deal of biasness was seen in that case. Though it was not the case for Random Forest because of its dependency on depth for prediction. But logistic regression depends on sigmoid and proper optimization of that would be good for our model.

-In logistic regression compared to the base and original model, a comparatively more score was achieved when modified data was used in our models. Feature reduction played a good role here in making the model more efficient. As in the case of logistic regression, the model depends on the sigmoid value, and changing the parameters or the features has an impact on it

Logistic Regression

According to (Kurt et al., 2008), logistic regression is useful for situations where one needs to predict the presence or absence of a feature or output, based on values of the set of variables to predict. It is similar to a regression linear model, but it is more appropriate for models where the dependent variable is dichotomic [8]

Pros

-Accuracy increases

-Bias decreases and also avoids overfitting

-Works well with non-linear data[7].

-Follows a rule-based approach, so data normalization is not required.

Cons

-Increases computational power

-Requires more time to compute

-Sometimes, failing to determine the significance of each variable

Hypothesis Statement

Two of the models selected are expected to perform well and give better accuracy in the result. Though our reference paper already dealt with logistic regression, still changing features, and doing some parameter changing can have an effect on the result.

Compared to the two models, RF is hypothesized to perform better in terms of accuracy prediction compared to the other model because. It can also be assumed that our modified dataset would have an impact in changing the accuracy of the model in case of LR. As few of our features are not normalised and as LR has affect while increasing or reducing the features, it can be assumed, LR model would perform good in modified dataset. Two of the models are very different from each other which would make this analysis and model implementation robust and interesting and many thoughtful insights can be got.

Random Forest Model Implementation

Parameter

-Hyperparameter optimization was carried out to find out the optimal parameters for Random Forest model. After 30 iterations, the best choice of parameters were-

Method = AdaBoostM2, NumLearningCycles = 462, MinLeafSize = 2

-The model was ran two times, one with default settings and the other time with the parameter values, the model performed well with the change and tweek of parameters and it was kept for the end. Experiments were carried out using two of the datasets on this model.

Result

Accuracy	95.54%
F1 Score	0.95

Logistic Regression Implementation

Parameter

-LR does not have too important parameters to tune[9]

-mnrfit() function was used with Model = hierarchical was passed as parameter

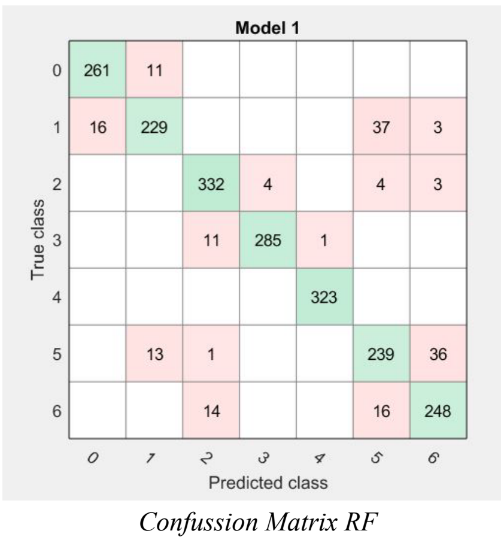
-Confusion Matrix was generated to give us precision and score

Result

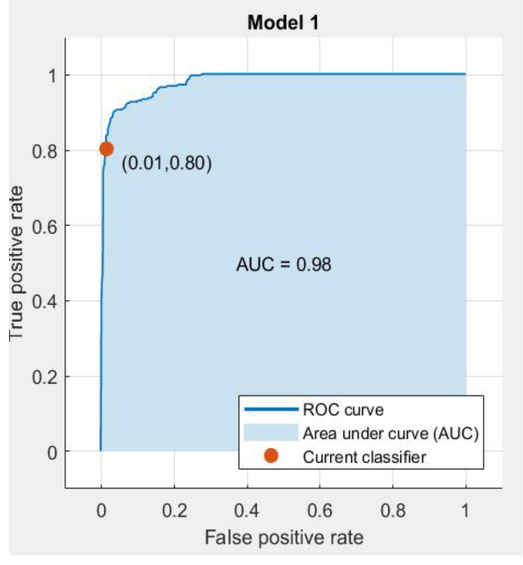
Accuracy	91%
F1 Score	0.55

Overall Results

	Normal Dataset	Dataset Feature Reduction and Standardization	Hyperparameter tuning
RF	Acc 91% F1 0.93	Acc 81.62% F1 0.81	Acc 95.54% F1 0.95
LR	Acc 78% F1 0.50	Acc 78% F1 0.55	



Confussion Matrix RF



ROC AUC LR

Lessons Learned and Future Work

-RF is highly customizable and robust model which gives good accuracy and scores by optimizing parameters and RF cannot properly handle feature reduction and normalization of data

-Time taken for RF to run the model was high compared to LR

-LR cannot deal with data which is imbalanced, accuracy increases by balancing the data and doing feature engineering and dimension and feature reduction

Future works- PCA should be applied to data and feature engineering should be done while using LR to get precise accuracy. One hot encoding can be done as part of feature engineering to see the results

References

[1]"https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=Of%20these%20over%20650%20million,overweight%20or%20obese%20in%202020.

[2]"Obesity and overweight," www.who.int. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=Of%20these%20over%20650%20million

[3] E. De-La-Hoz-Correa, F. E. Mendoza-Palechor, A. De-La-Hoz-Manotas, R. C. Morales-Ortega, and S. H. Beatriz Adriana, "Obesity Level Estimation Software based on Decision Trees," Journal of Computer Science, vol. 15, no. 1, pp. 67–77, Jan. 2019, doi: 10.3844/jcssp.2019.67.77.

[4] Europe PMC, "Europe PMC," Europepmc.org, 2019. https://europepmc.org/article/pmc/pmc4586339

[5] C. Aday et al., "Machine Learning Approaches for the Prediction of Obesity using Publicly Available Genetic Profiles." [Online]. Available: https://researchonline.ljmu.ac.uk/id/eprint/5450/9/Machine%20Learning%20Approaches%20for%20the%20Prediction%20of%20Obesity%20using%20Publicly%20available%20genetic%20profiles.pdf

[6] P. S. Kumar and S. Pranavi, "Performance analysis of machine learning algorithms on diabetes dataset using big data analytics," IEEE Xplore, Dec. 01, 2017. https://ieeexplore.ieee.org/document/8286062

[7] T. A. Team and T. A. Team, "Why Choose Random Forest and Not Decision Trees – Towards AI — The Best of Tech, Science, and Engineering." https://towardsai.net/p/machine-learning/why-choose-random-forest-and-not-decision-trees

[8] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," Expert Systems with Applications, vol. 34, no. 1, pp. 366–374, Jan. 2008, doi: 10.1016/j.eswa.2006.09.004.

[9] J. Brownlee, "Tune Hyperparameters for Classification Machine Learning Algorithms," Machine Learning Mastery, Dec. 12, 2019 https://machinelearningmastery.com/hyperparameters-for-classi-