Predicting obesity levels based on eating habits and physical condition

Mansion See Lui

April 3, 2022

Abstract

In this paper, I present a machine learning based approach to projecting the obesity levels based on eating habits and physical. With the proliferation of data, analytics have increasingly become a critical component in the assessment of a person obesity level. I used the data sets from UCI.

1 Introduction

Obesity is a complex, multifactorial, and largely preventable disease, affecting, along with overweight, over a third of the world's population today. If secular trends continue, by 2030 an estimated 38 percent of the world's adult population will be overweight and another 20 percent will be obese.

Obesity is typically defined quite simply as excess body weight for height, but this simple definition belies an etiologically complex phenotype primarily associated with excess adiposity, or body fatness, that can manifest metabolically and not just in terms of body size. Obesity greatly increases risk of chronic disease morbidity—namely disability, depression, type 2 diabetes, cardiovascular disease, certain cancers—and mortality. Childhood obesity results in the same conditions, with premature onset, or with greater likelihood in adulthood.

In this paper, the dataset from the UCI will be use to predict the level of obesity of a person. Note that the dataset is very specific to individuals from the countries of Mexico, Peru and Colombia.

2 Dataset Selection

2.1 Dataset Overview

The data is predicated on our ability to design a model that could predict the obesity level of a person. The data for the estimation of obesity levels in people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition. Data was collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 17 attributes and 2111 records. The attributes related with eating habits are: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS), other variables obtained were: Gender, Age, Height and Weight. Finally, all data was labeled and the class variable NObesity was created with the values of: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. The data contains categorical data and continous data, so it can be used for analysis based on algorithms of classification.

3 Exploratory Data Analaysis

3.1 Pre-Modelling/Data Cleaning

The data as mention consist categorical and continuos data. Those categorical data was change into numerical data. Subsequent checking of individual data after changing to numerical data and small amount of data spread need to be clean up.

Two spread of data was decided to clean up. First, feature CALC has 4 category and one of the category has only one entry. I decided to remove the row of data. Reason is, the data is not enough for the machine to learn and predict from it. Second, feature MTRANS has 5 category but 2 of the category also faces a few number only. I decided it to merge it to other category in order to have a significant data set. The category change for MTRANS are bike and motorbike, and was merge into category Walking. So Walking/Motorbike/Bike form one category as a results.

The classification of the data spread is quite balance. The python file will show all the above Pre-Modelling/Data clean up process.

4 Model Determination

4.1 Model Types

I try to setup a few multi class classification model to solve the problem. The model I have tried for classification are: 1. Logistic Regression 2. Logistic Regression CV (solver: Newton-cg) 3. Logistic Regression CV (Solver: lbfgs) 4. Decision Tree 5. Random Forest 6. Extra Tree 7. Support Vector Machine 8. Naive Bayes (Gaussian) 9. Stochastic Gradient Descent

4.2 Model Selection Criteria

As stated above, we will be testing different classification algorithms in an attempt to find the best fit for our model predictions. The metrics of Accuracy, Confusion Table and AUC will determine which algorithm best for the datasets.

5 Analysis

5.1 Comparing Model I

As mention, different type of model was used. Below figure will show the results of each model type.

	lr	l2_lbfgs	l2_newton-cg	dt	rf	gnb	sgd	suvm
precision	0.733781	0.964356	0.966890	0.919396	0.952751	0.640821	0.707889	0.938833
recall	0.740476	0.964286	0.966667	0.916667	0.947619	0.604762	0.692857	0.938095
fscore	0.732253	0.964045	0.966446	0.917097	0.948083	0.583812	0.689888	0.937827
accuracy	0.740476	0.964286	0.966667	0.916667	0.947619	0.604762	0.692857	0.938095
auc	0.848482	0.979211	0.980616	0.951450	0.969465	0.767626	0.821190	0.963982

Figure 1: Comparison Between each Model

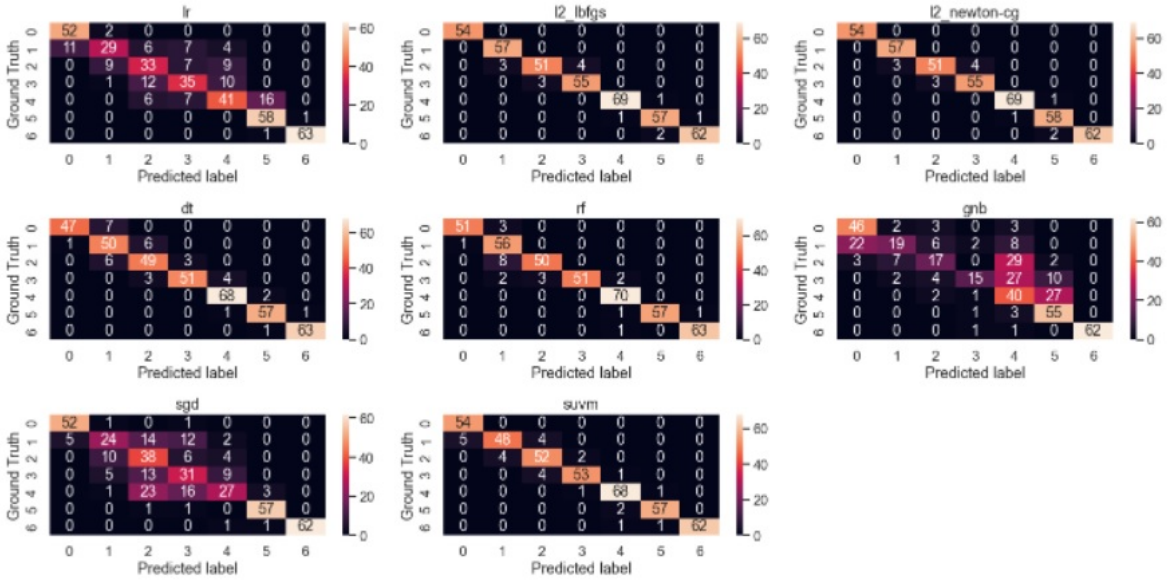

Figure 2: Confusion Matrix

5.2 Model Selection I

Based on the results shown in Figure 1 and Figure 2 and based on the metrics mention, there are five model that are quite close to each other. 1. Logistic Regression CV (solver: Newton-cg) 2. Logistic Regression CV (Solver: lbfgs) 3. Decision Tree 4. Random Forest 5. Support Vector Machine

5.3 Comparing Model II

Base on the category we are predicting which is the level of obesity, it is self explanatory that the weight will play a big factor. So, I decided to drop the weight as feature and further analyze the results using the selected model at Model Selection I. Below will show the results:

	l2_lbfgs	l2_newton-cg	dt	rf	suvm
precision	0.572389	0.573608	0.774621	0.880833	0.637633
recall	0.576190	0.576190	0.771429	0.869048	0.638095
fscore	0.564215	0.564750	0.771039	0.870982	0.630123
accuracy	0.576190	0.576190	0.771429	0.869048	0.638095
auc	0.752236	0.752275	0.866624	0.923933	0.789248

Figure 3: Comparison Between each Model

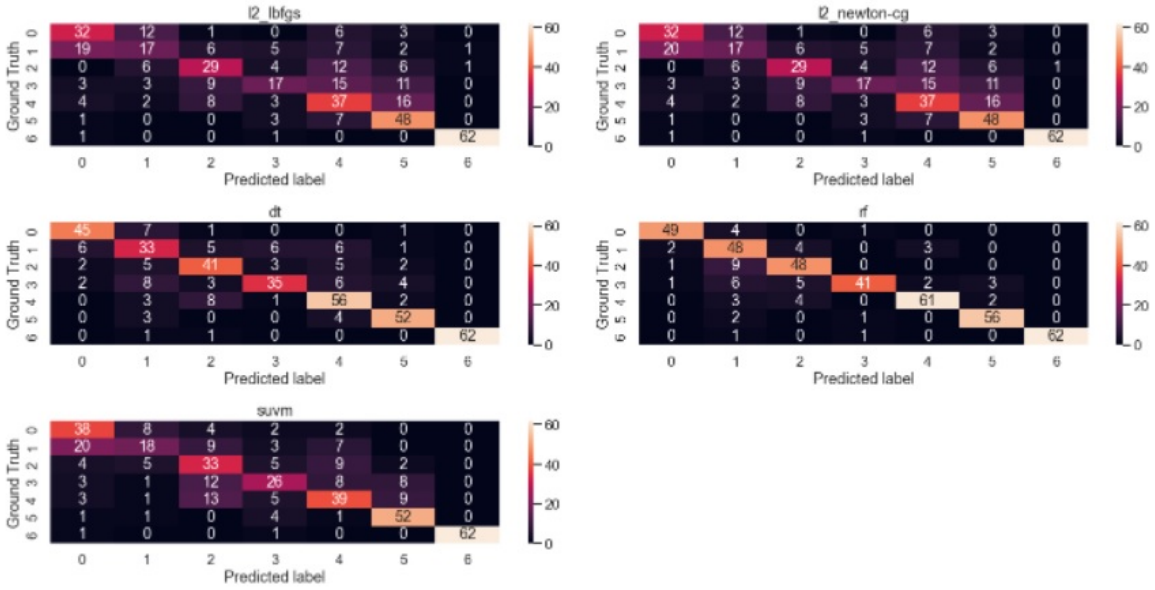

Figure 4: Confusion Matrix

5.4 Model Selection II

Based on the results shown in Figure 3 and Figure 4 and based on the metrics mention, the Random Forest shows a better performance with and without the feature weight.

5.5 Supporting Analysis

I used the Feature Importance to show why when we drop the feature weight, the other algorithm are not performing well.

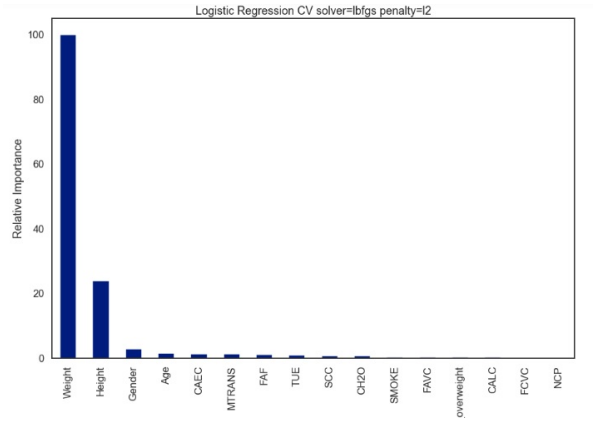

Figure 5: Feature Importance Logistic Regression CV

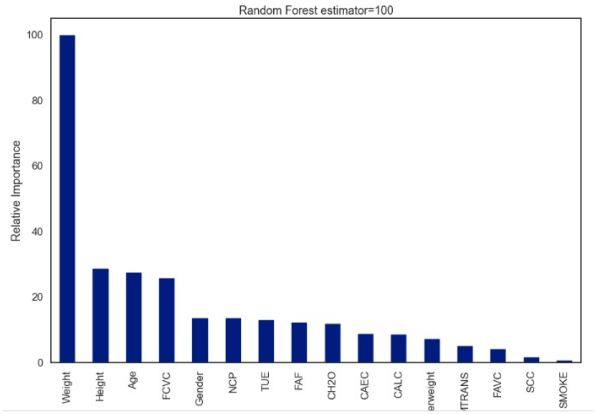

Figure 6: Feature Importance Random Forest

It shows that the feature weight for logistic regression is very much depend on it. So without the feature weight, the performance will not do well. Unlike for random forest, the weight is also dominant but other factors also like height, age and others can play a factor. So when we remove the feature weight and compare those algorithm, the random forest shows a better performance. Below figure show the random forest feature importance without feature weight.

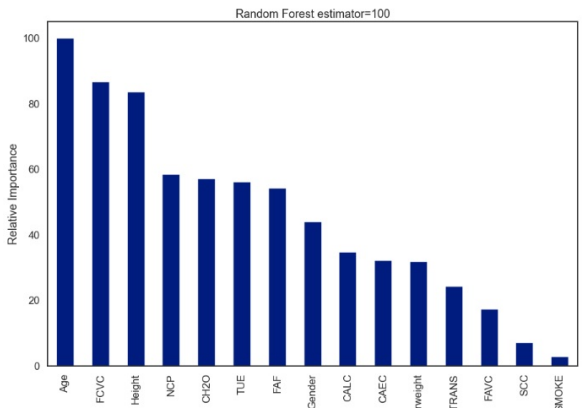

Figure 7: Feature Importance Random Forest without feature weight

5.6 Full Model

With random forest as the model, I try to find the best hyperparameter to fit into the model. The python file will have all the detail of the results. After comparing the best parameter in Grid Search and the original parameter, the difference is very small and it seems that the original parameter performs a little better. The full model below:

RandomForestClassifier(oob_score=True, random_state=42, n_jobs=-1, bootstrap=True, class_weight=None, criterion='gini', n_estimators=100)

6 Conclusions

With so many type of classifier to use to classify a category, we should choose a model that will be able to classify at a higher percentage with or without the main feature importance. In this case, the feature weight is the main feature across all model. But when the feature weight was taken out, some model perform much better than other. For this problem, we can say that random forest will be the better choice as compare to the other model based on the accuracy, confusion matrix and AUC with and without weight as features.