Estimation of obesity levels based on eating habits and physical conditions.

Jie Bai [a], Carolina Ramirez-Tamayo [b]

Department of Mechanical Engineering, University of Texas at San Antonio, San Antonio, Texas

jie.bai@utsa.edu [a] , caro0613@hotmail.com [b]

PhD. Adel Alaeddini.

Group 6.

Estimation of obesity levels based on eating habits and physical conditions.

This report presents the development of a Sequential Model, used to predict obesity levels in individuals from Colombia, Mexico, and Peru, based on a dataset provided by UCI Machine Learning repository, that contains the data obesity levels based on their eating habits and physical conditions.

The performance of the model is desirable compared with previous studies using the same dataset, with a training accuracy of 99%, and test accuracy of 96%, and can be used to predict obesity levels.

Keywords: machine learning, artificial neural networks, accuracy, habits, obesity.

Introduction

Obesity has become more and more epidemic in every country of the world that leads to serious consequences for people of all ages. It has been shown that obesity can be considered a disease that mainly relates to multiple factors such as food type consumption, physical conditions, and family history of obesity, etc. Obesity has drawn great attention from researchers to investigate the influence factors of obesity and at the same time to predict the emergence of obesity under these factors by using advanced technological tools such as machine learning.
In this group project, we were able to access the “Estimation of obesity levels based on eating habits and physical condition Data Set” (Palechor and Manotas) that include the structured data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia based on their eating habits and physical conditions to study on how to predict the obesity level based on the factors of eating habits and physical conditions by implementing the deep learning algorithm on the data set. First, we did some exploration and statistical analysis on the raw data. The data showed different types of obesity between different gender male and females. The obesity categories are also different among different age groups, height groups and weight groups. Then necessary data preprocessing including imputation on the missing values, normalization, scaling, encoding have been performed to make the data ready to be trained by machine learning algorithms. Finally, the data was trained by a 3-layer Artificial Neural

Networks (ANN). With the ANN model we developed in this project, 96% accuracy was achieved, which is a good performance model compared with the results of the previous prediction models of Naïve Bayes and simple logistic regression (De-La-Hoz-Correa et al.).

Literature Review

Obesity has drawn research’s interest nowadays, and many studies are now focusing on the investigation of the factors that lead to obesity disease. In this section, we are briefly reviewing the previous works presented using different techniques on datasets related to the obesity health issue. In 2009, Zhang et al.(Zhang et al.) presented a logistic regression model to predict the overweight for children at different age groups, and compared its model with several other techniques including Decision Tree, Association Rules, Neural Networks, Naïve Bayes, etc. In 2012, a framework based on Naïve Bayes with a hybrid approach and genetic algorithms was presented by Adnan and Husain (Adnan and Husain) to predict children's obesity. There are 19 parameters in prediction and results in a 75% precision. Then Husain et al. (Husain) built an intervention system for primary schools aiming at reducing children obesity in 2013, the program was called MyHealthyKids, and it was based on Naïve Byes to identify the obesity probability for children. It showed a 73.3% precision by tests. The producer of the dataset we are using in this project developed an obesity level estimation software based on decision trees (J48) in 2019, which yield a high precision rate of 97.4% (De-La-Hoz-Correa et al.). In this project study, we are implementing the artificial neural networks we learned from the Deep Learning class on the same dataset used in the work done by De-La-Hoz-Correa et al. in 2019, to see if our ANN model can achieve good performance on the prediction of obesity levels.

Approach

In the given dataset, there are 17 attributes including the attributes related with eating habits, physical conditions, and some other variables to be considered to estimate the obesity levels with 7 categories. Among the 17 input features, some of them are numerical features while the others are categorical features. Since our features are not sequenced data, we used mean imputation to substitute the missing values of the numerical features with the mean value of that feature. Also, normalization needs to be implemented to our data since different features have different scales. Here we normalize our data with standard score that takes the formula: $X'=X-\mu$, where is the mean of the samples and is the standard deviation of the samples. Since classifiers cannot operate with categorical data directly, one hot encoder and label encoding is used to assign numerical values to each category. One hot encoding creates a binary column for each category and returns a sparse matrix or dense array. There is a problem we must deal with is after all the categorical features have been encoded, the total number of features increased. Having too many features, especially having irrelevant features in our data can decrease the accuracy of the models. Therefore, we must reduce the dimensionality. In this project, we reduce the dimensionality by doing feature selection with tree-based method. To perform the prediction, we use the artificial neural networks (ANN) as our baseline model since it is being used widely in healthcare as predictive models as it achieves high accuracy results when dealing with larger training datasets at great speeds, and it have applications in nearly all departments of a hospital, especially in the fields such as cancer diagnosis and cardiac arrhythmias. In ANN, the neurons are typically organized into multiple layers. Neurons of on layer connect only to neurons of the immediately preceding and immediately following layers. The training data are inputted in the first layer called input layer. The layer that produces the ultimate result is the output layer. And there are one or multiple hidden layers between the input and output layers, where a group of neurons in one layer connect to a single neuron in the next layer. The architecture of ANN is represented in Fig.1.

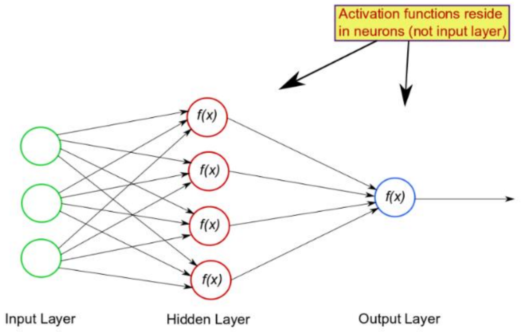


Figure 1: Architecture of Artificial Neural Network

To find the output of the neuron, first we take the weighted sum of all the inputs, weighted by the weights of the connections from the inputs to the neuron. We add a bias term to this sum. This weighted sum is then passed through a (usually nonlinear) activation function to produce the output. The procedure can be described by the following formula:

$$Z(f(x) = W_i x_i + b_i \qquad (1)$$

$$\mathrm{ReLU}(z)=\max(0,\ z) \qquad (2)$$

Since in our case it is a supervised learning, mean-square-error is used as the cost. The baseline model can be improved by lowering the learning rate, optimizing the ANN structure, or training the model using gradient descent with back propagation to minimize the squared error.

Experimentation

1) Data:

The data was collected from UCI Machine Learning Repository. “The data contains 17 attributes and 2111 records, the records are labeled with the class variable NObesity

(Obesity Level), that allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform. This data can be used to generate intelligent computational tools to identify the obesity level of an individual and to build recommender systems that monitor obesity levels. For discussion and more information of the dataset creation, please refer to the full-length article "Obesity Level Estimation Software based on Decision Trees" (De-La-Hoz-Correa et al., 2019)". (Fabio Mendoza Palechor)

2) *Evaluation Model:*

Considering that our model is non-linear and has several inputs and outputs, we decided to use Artificial Neural Networks, sequential model. After 500 epochs, using as optimizer the "Adam Optimizer", changing the learning rate parameter, our evaluation method was testing the accuracy of the model.

We divided the evaluation of the model in two parts: a) testing the accuracy of the training data, b) testing the accuracy of the test data. The results were pretty accurate, 99% and 96% of accuracy, respectively.

3) *Experimentation Detail:*

For the experiment detail, we developed it in several parts, as follows:

a) Downloading the dataset and reading about it was the first step, it gave us a general idea of what we were about to work in.

b) Importing the libraries, we needed to develop the code.

c) Data Exploration. In this section we explored our dataset, trying to find patterns, understanding the behaviour of the dataset, plotting the parameters, finding qualitative conclusions and relation between parameters.

d) Encoding our Data. Since this is a classification task, classifiers cannot operate with categorical data, the use of hot encoders and label encoders was needed, to transform the qualitative data into numbers. The result of encoding was having more features, that later were labelled, scaled, and normalized.

e) Feature selection. In previous step, the result was having more features, but having too many features can decrease the accuracy of the model, so this task was focused on identifying which features were relevant in our model.

As the follow graph shows, some features were irrelevant and others, like weight and height, represents a big portion if the model's importance. The features we decided to work with, are: Weight, Age, Height, Frequency of consumption of vegetables are the most 4 important features.

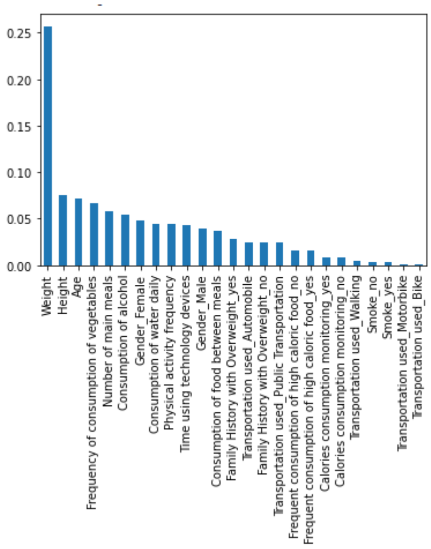


Figure 2: Feature Selection

f) Model Selection. In this section, we previously were assigned to work either with ANN (Artificial Neural Networks) or LSTM (Long Short-Term Memory). We decided to work with Artificial Neural Networks, sequential model, because our model is expected to have a categorical output (7 outputs), several inputs (features), and it allows us to continuously change the hyperparameters (learning rate, optimizer, metrics, and loss).

In general terms, the model performs pretty accurate with our initial hypothesis: building a model with high accuracy to predict obesity levels in individuals, according to their habits. We present our results in the next section.

Discussion/Analysis

For the discussion/results, we divided it into two parts: Qualitative and Quantitative results.

1) Qualitative Results:

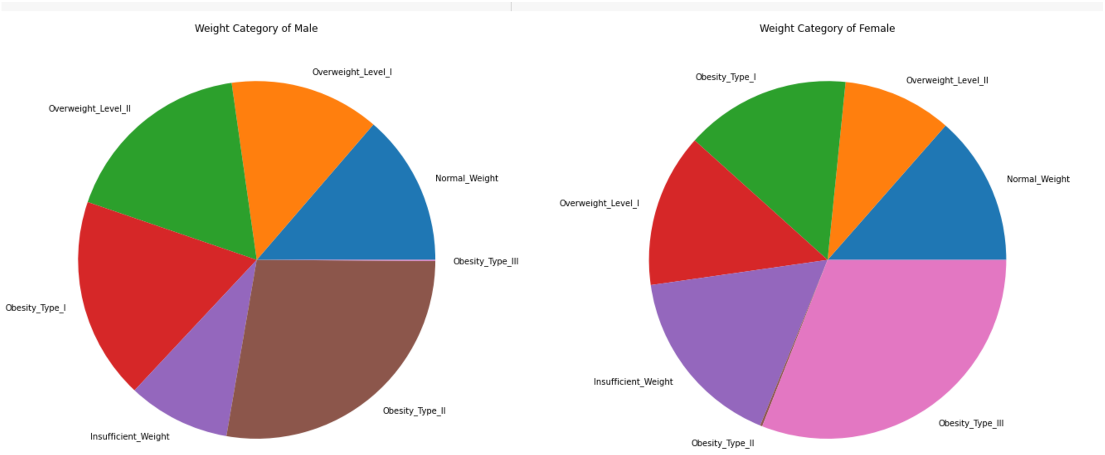


Figure 3: Qualitative Results

A bigger proportion of female with a large slice of Obesity Type III in the pie chart below, while Obesity Type II is the most prevalent type of obesity in male. Interestingly, there is also a higher proportion of Insufficient Weight in female compared to male, this could be explained by a heavier societal pressure on women to go on diets.

```
<AxesSubplot:>
```

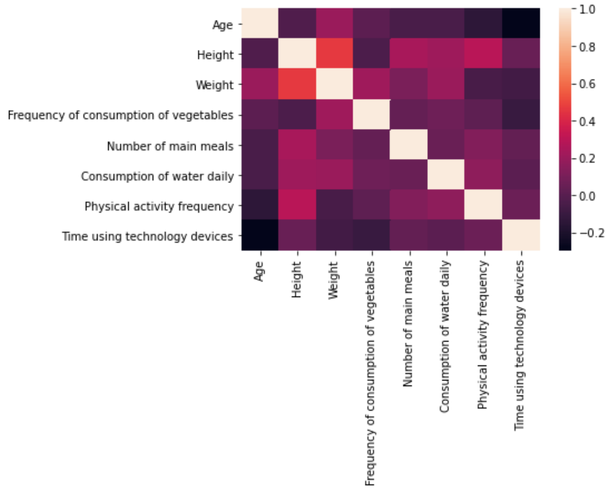


Figure 4: Correlation Matrix

There is a direct relation between Height and Weight, which can be seen when we performed the feature selection.

2) Quantitative Results:

After running the 3-layer ANN model with a learning rate of 0.0001 for 500 epochs, we plotted the train/test loss and train/test accuracy in figure 3 and figure 4, respectively. From the two plots, both the training loss and the testing loss were decreasing as more epochs were run. Same behaviour can be seen with the plot of accuracy. The test loss is 0.093 which is higher than the 0.0024 training loss, and the train accuracy is 0.99 which is lower than the 0.96 test accuracy. Also, we can see from figure 1 that our test accuracy is smaller than the train accuracy which indicates our model is not overfitting. There are7 obesity categories in the output, with each category yielding a more than 0.9 f1-score, as shown in table.1. Both precision and recall of each category are desirable. Our model obtained 96% accuracy on prediction, higher than 90% precision levels to classify people that carry a level of obesity for almost all 7 obesity levels except for only one level have 86% precision. Same level of recall score was also obtained that indicated our model was able to find all the relevant cases within the dataset. The prediction of obesity levels using artificial neural networks can be implemented in the intervention of obesity related diseases.

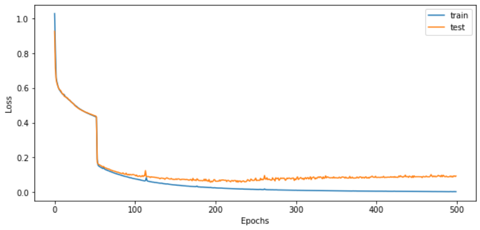


Figure 6: Train/Test Loss

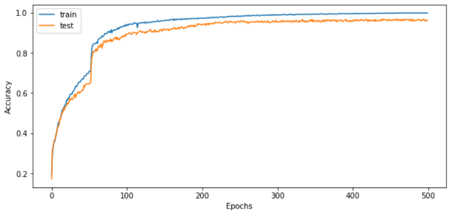
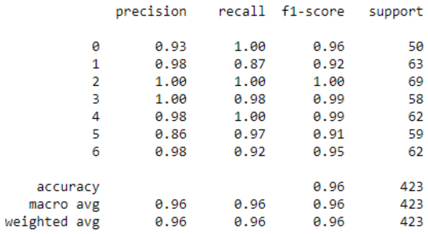


Figure 5: Train/Test Accuracy

Table 1: Error Analysis



```
              precision    recall  f1-score   support

           0       0.93      1.00      0.96        50
           1       0.98      0.87      0.92        63
           2       1.00      1.00      1.00        69
           3       1.00      0.98      0.99        58
           4       0.98      1.00      0.99        62
           5       0.86      0.97      0.91        59
           6       0.98      0.92      0.95        62

    accuracy                           0.96       423
   macro avg       0.96      0.96      0.96       423
weighted avg       0.96      0.96      0.96       423
```

Conclusion

To summarize, in this project we implemented a 3-layer artificial neural network on obesity level prediction based on multiple attributes such as people's eating habits and physical conditions. Our model obtained 96% accuracy on prediction, higher than 90% precision levels to classify people that carry a level of obesity for almost all 7 obesity levels except for only one level have 86% precision. Same level of recall score was also obtained that indicated our model was able to find all the relevant cases within the dataset. The model developed in this study allows us to successfully classify patients with obesity and might contribute to the medical diagnosis and intervention in obesity related diseases.

References

Adnan, Muhamad Hariz B Muhamad and Wahidah Husain. "A Hybrid Approach Using Naïve Bayes and Genetic Algorithm for Childhood Obesity Prediction." 2012 International Conference on Computer & Information Science (ICCIS), vol. 1, IEEE, 2012, pp. 281-285.

Davila-Payan, Carlo. "Estimating Prevalence of Overweight or Obese Children and Adolescents in Small Geographic Areas Using Publicly Available Data." Preventing chronic disease, vol. 12, 2015.

De-La-Hoz-Correa, Eduardo et al. "Obesity Level Estimation Software Based on Decision Trees." 2019.

Husain, Wahidah, M.H.M. Adnan, L.K. Ping, J.Poh and L.K. Meng. "My Healthy Kids: Intelligent Obesityintervention System for Primary School Children." Proceedings of the 3rd International Conference on Digital Information Processing and Communications, (IPC' 13)

vol. The Society of Digital Information and Wireless Communication, 2013, pp. 627-633.

OMS. WHO Official Website, 2021.

Palechor, Fabio Mendoza and Alexis de la Hoz Manotas. "Dataset for Estimation of Obesity Levels Based on Eating Habits and Physical Condition in Individuals from Colombia, Peru and Mexico." Data in Brief, vol. 25, 2019, p. 104344, doi: https://doi.org/10.1016/j.dib.2019.104344.

Zhang, Shaoyan et al. "Comparing Data Mining Methods with Logistic Regression in Childhood Obesity Prediction." Information Systems Frontiers, vol. 11, no. 4, 2009, pp. 449-460.