Machine-Learning-Based Prediction of Obesity

Siyuan Chen[a]
International Business School, Henan University, Zhengzhou, China

Keywords: Machine Learning, Obesity Prediction, Neural Network.

Abstract: Obesity is a common phenomenon today. It is a chronic metabolic disease caused by excessive fat accumulation and is the result of the interaction of multiple factors such as genetics and environment. As machine learning is widely used in various fields, obesity data is processed by using machine learning methods to obtain fitting models, to realize the prediction of obesity and to determine the main causes of obesity. The main contents of this paper include: (1) using the obesity data provided by the UCI data set as the research object, using a series of preprocessing data in Python language. (2) Establish a machine learning model and import the data to generate bar charts and other graphs that reflect the relevant results. (3) Results analysis, to Determine which factors are more closely related to obesity and evaluate the performances of the models. The comprehensive analysis of accuracy, recall rate, precision rate and other indicators finally obtains the best prediction effect of the GBDT algorithm, which can effectively predict obesity.

1 INTRODUCTION

Nowadays, people's living standards are constantly improving with the development of society, and their diet is becoming more and more diversified. While enjoying these conveniences, there are also some potential risks. The pace of society is accelerating, and people are under pressure from all sides. With this situation, people's health constantly produces various problems. Obesity is a common phenomenon, it is caused by genetic and environmental factors such as chronic metabolic diseases, the Chinese Residents Nutrition and Chronic Condition Report (2020), according to more than half of adults overweight / obesity, 6~17, children under 6 and adolescents overweight / obesity rate reached 19.0% and 10.4% respectively (Liu, 2021). Obesity not only affects the external appearance but also leads to various other diseases. At the same time, to prevent and treat obesity, people need to spend more time and money. Therefore, the prevention and treatment of obesity is of great significance to prevent chronic diseases and reduce the personal financial burden. With artificial intelligence developing so quickly, machine learning has become widely applied in many different fields, the medical field is also a big aspect of machine learning application, through the use of the machine learning function of existing data analysis, which can conclude that the relevant data about obesity to determine what are the main causes of obesity.

From the point of the current development situation, although there have been some machine learning methods applied to disease prediction, for the prediction of obesity is relatively lack of comprehensive and perfect research, so this paper based on machine learning from a variety of algorithms to find out the good prediction of obesity prediction model, to realize the timely management and treatment of obesity.

2 METHOD

This chapter introduces six machine learning models of logical regression, decision tree, random forest, GBDT, XGBoost and deep learning network (DNN), and presents the experimental data set and the unbalanced learning methods, and expounds the main content and direction of this experiment.

2.1 Logistic Regression

Among the linear regression models is logistic regression, which uses data from independent

[a] https://orcid.org/0009-0006-3168-8559

Chen, S.
Machine-Learning-Based Prediction of Obesity.
DOI: 10.5220/0012916000004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 153-158
ISBN: 978-989-758-713-9

variables as input to forecast the chance of a desired result. Logistic regression has similarities to the principle of multiple linear regression, which first determines the best-fitting regression line to represent the connection between the independent variable (x) and the dependent variable (y), the regression line (Wang, 2022). The model form is related to the linear equation. If the independent variable is a data set, the equation can be shown as a matrix: $y = ax + b$

$$z = a^{(1)} \cdot x^{(1)} + a^{(2)} \cdot x^{(2)} + a^{(3)} \cdot x^{(3)} + \cdots + a^{(n)} \cdot x^{(n)} + b \quad (1)$$

Logistic regression functions the linear equation corresponding to a state p, determining the size of the dependent variable based on the values of p and 1-p. The dependent variables of logistic regression can be dichotomous or multiple classifications, and the independent variables can be continuous or discrete.

There are three types of logistic regression: ordinal, multinomial, and binary.

The results of this paper are divided into obesity and non-obesity after data processing, so binomial logistic regression is adopted. The dependent variable of binomial logistic regression is essentially a dichotomy method, that is, there are only 0 or 1 results, and the probability distribution is as follows:

$$P(Y=1 \mid x) = \frac{e^{w^T \cdot x}}{1 + e^{w^T \cdot x}}, \quad P(Y=0 \mid x) = \frac{1}{1 + e^{w^T \cdot x}} \qquad (2)$$

2.2 Decision Tree

The decision tree technique is a prediction technique for creating target variables or a categorization scheme based on several variables. This algorithm can effectively handle large data sets. Common uses of the decision tree model include variable selection, evaluation of variable importance, processing of missing values, and prediction (Song, 2015). The duality of the results enables a good application of decision trees for obesity prediction in this experimental study.

The main parts of the decision tree model are the nodes and branches and the construction of the model includes splitting, stopping and pruning (Song, 2015).

The nodes of the decision tree can be divided into three types: (1) root nodes, also called decision nodes. (2) Internal nodes. (3) The decision tree's ultimate outcome is represented by the leaf node, sometimes referred to as the end node.

The decision tree is a continuous model that combines a series of tests and compares the feature values in each test to the threshold value(Navada, Ansari, Patil, 2011). Each node in the decision tree corresponds to an analysis of data properties, The decision tree model links the dataset's observations to the conclusions(Sharma, Kumar, 2016) of the pertinent target values, with each branch denoting the analysis's findings.

2.3 Random Forest

The classification and regression tree model is further improved by the random forest technique, which is composed of a large number of decision trees created by randomization, which can be used for prediction once constructed. Because the validity of the decision tree for binary classification applies to the prediction of obesity, the random forest model was adopted by this experiment. The average of the outputs of a random forest with several decision trees is aggregated into a single output of reference (Rigatti, 2017). Formally, a model made up of several random base regression trees is called a random forest {rn (x, Θ m, Dn), m 1}, where Θ 1, Θ 2,... Is the output of the random variable, Θ. Combining these random trees forms an aggregate regression estimate of research (Biau, 2016). The formula is as follows:

$$\overline{\gamma}_n(X, D_n) = \mathrm{E}_\Theta[\gamma_n(X, \Theta, D_n)] \qquad (3)$$

Based on each decision tree produces a result based on the input data; and the final output of the random forest is the paper obtained after integrating the multiple results (Liu, 2014).

The advantage of random forest is that it can find the interaction between the predicted variables and the non-linear relationship, but it is difficult to judge which variables have a greater impact on the prediction results.

2.4 GBDT Model

GBDT, whose full name is a gradient-lifting decision tree, is an iterative decision tree algorithm that is also applicable to the binary characteristics of this study. To obtain the ultimate prediction outcome, the model aggregates the outcomes of several decision trees. The next weak classifier fits the residual function of the predicted value, which is the difference between the predicted value and the true value, and the principle adds the results of all the weak classifiers to the predicted value. The decision tree is the common learner in the GBDT model, which is an integrated learning model (Zhang, 2021).

2.5 XGBoost Model

XGBoost is an improved model of the gradient boosting algorithm and a machine learning algorithm implemented under the Gradient Boosting framework.

The basic component of the XGBoost model is the decision tree, which has good results for dichotomous data analysis, so it is applied to the prediction of obesity in this experiment. There is an order between them: the latter decision tree will be combined with the prediction results of the previous decision tree, that is, take the analysis error of the previous tree into account, which increases the proportion of the error between the previous samples in the subsequent prediction, thus improving the accuracy and scientificity of the model prediction. According to the model principle, the prediction calculation formula is obtained as follows:

$$\hat{y}_t^{(a)} = \hat{y}_t^{(a-1)} + f_a(x_t) \qquad (4)$$

That is the predicted value of the first tree for the sample t = the predicted value of the first a-1 tree for the first t-1 sample + the first tree for the sample t. The objective function of the model prediction result is obtained as follows:

$$I^{(a)} = \sum_{t=1}^{n} l(y_t, \hat{y}_t^{(a)}) + \Omega(f_a) \qquad (5)$$

2.6 Deep Neural Network

The DNN is designed to predict the data set. The model construction of the neural network includes input, output and three hidden layers. Through the excitation function tanh () in the hidden layer, the nonlinearity of the data set can expand the expression ability of the neural network (Figure 1). The tanh function is centered on zero, the gradient is steeper and not limited to one direction, and it is overall superior to the sigmoid function (Sharma, 2017).

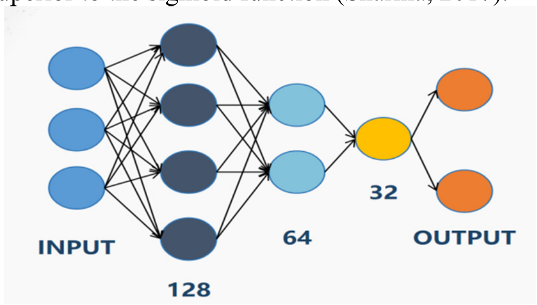

Figure 1: Deep Neural network model (Photo/Picture credit : Original).

Deep neural networks have the good nonlinear fitting ability but require large-scale datasets for training, otherwise overfitting may occur.

3 RESULTS AND DISCUSSION

3.1 Pro-Processing

3.1.1 Distribution of the Data Sets

As shown in pie figure 2, the proportion of obesity in the surveyed subjects in this experimental dataset is not balanced compared with other resultPs, so it is speculated in the dataset.

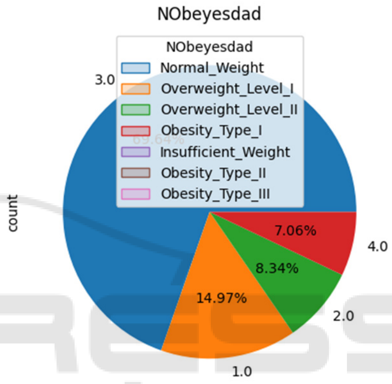

Figure 2: Schematic representation of the obesity ratio among the survey subjects (Photo/Picture credit : Original).

3.1.2 Unbalanced Processing Algorithm

The SMOTE algorithm, namely the Synthetic Minority Oversampling Technique, also known as the synthetic minority oversampling technology, is also an improved random oversampling method that generates new artificial minority instances by interpolating in several instances located together (Yang, 2021). It is based on sampling data from a few classes by connecting random data points (Elreedy, 2019). SMOTE can alleviate the overfitting phenomenon of random oversampling by inserting synthetic instances in a new position, but it still has two disadvantages, one is that it can spread the noise, and the other is that in the SMOTE algorithm, all instances share the same global neighborhood parameters, which will ignore the distribution characteristics, resulting in poor classification effect. The SMOTE algorithm is very robust, but in the case of chaotic sample data and some scattered samples, it will mechanically generate new points, which will become noise points and affect the classification performance of the model (Yang, 2021).

3.1.3 Hyperparameter Setting

The experiment code was run using Python language through Pycharm compilation software. Where the main hyperparameter package Epochs, batch_size.

Epochs refer to the number of times the training set is fully trained when training the neural network, also called the process of neural network forward propagation and backpropagation.

The amount of samples that the neural network chooses to use in a single training session is known as the batch_size.

In this experiment, the results of different parameters were analyzed for multiple times, so that the set epochs value is 200 and the base _ size value is 70, and the comparison results between models obtained with the above values are the most balanced.

3.2 Evaluation Indicators

(1) Accuracy

Model accuracy (Accuracy) is one of the main indicators to evaluate the model performance, and the following is its calculating formula:

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

(2) Precision Rate

Accuracy is the proportion of samples with positive model prediction results in all positive samples. The following is the calculating formula:

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP} \tag{7}$$

(3) Recall

The percentage of samples that the model predicts is referred to as the recall rate. The following is the calculating formula:

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{8}$$

When evaluating the model with recall rate, a model's recall rate is low; when the precision rate is low, its recall rate is relatively high.

(4) F1-score

The F1-score, also known as the F1 score, is an indicator that combines precision and recall rate to produce average results. The formula is:

$$\mathrm{F1} = \frac{2 \times \mathrm{Precisi} \times \mathrm{Recall}}{\mathrm{Precisi} + \mathrm{Recall}} \tag{9}$$

Using the F1 score can avoid the problem that the data value is relatively close and it is difficult to choose and can find the model with better performance more efficiently.

(5) AUC

The meaning of AUC is the area under the curve and the axis, here the curve is the ROC curve. The area under the curve can be more intuitively judged by the model performance.

3.3 Analysis of the Indicators

(1) Precision (accuracy)

First, observe the prediction effect of each model on the data set through accuracy, and the comparison figure is as follows (Figure 3 and Table 1):

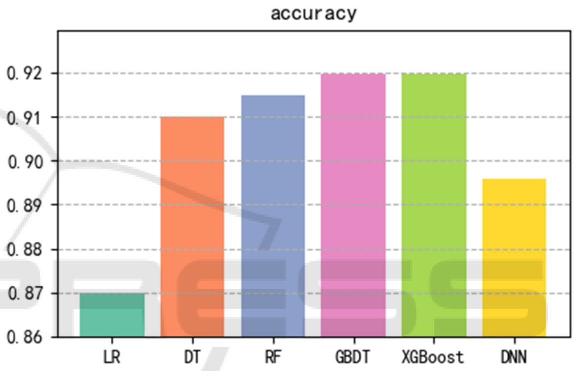

Figure 3: The prediction accuracy of each model (Photo/Picture credit : Original).

Table 1: The prediction accuracy of each model.

Method	accuracy
Logistic Regression	0.87
Decision Tree	0.91
Random Forest	0.915
GBDT	0.919
XGBoost	0.92
DNN	0.897

As illustrated in Figure 3, the logistic regression model has the lowest prediction accuracy, only around 87%, followed by DNN, and the models with higher accuracy were GBDT and XGBoost.

Because the result of the data prediction is binary, the model based on the tree model prediction accuracy is high, the GBDT in predicting new samples, each tree will produce an output value, the output value superposition, to get the final prediction value, for each tree training is the difference is the true result of the next tree prediction, for the experimental data set its data difference is large, so the test GBDT model prediction accuracy is high. The XGBoost model is similar to the GBDT principle, and the result of the previous tree during training affects the generation of the latter tree, so the prediction accuracy of this data set is relatively high.

(2) Accuracy Rate (precision) and Recall Rate (recall) The precision and recall of each experimental data set are shown in the figure below (Figure 4 and Table 2 Table 3):

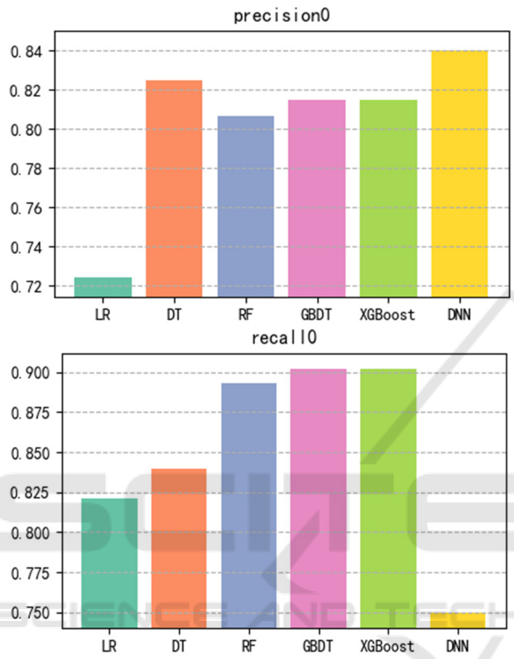

Figure 4: Prediction accuracy and recall rate of each model(Photo/Picture credit : Original).

Table 2: Prediction accuracy of each model.

Method	Precision0
Logistic Regression	0.723
Decision Tree	0.822
Random Forest	0.806
GBDT	0.817
XGBoost	0.818
DNN	0.84

Table 3: Recall rate of each model.

Method	Recall0
Logistic Regression	0.824
Decision Tree	0.84
Random Forest	0.89
GBDT	0.905
XGBoost	0.904
DNN	0.75

As demonstrated in Figure 4, the relationship between precision and recall rate. Among them, the comparison of the DNN model is the most obvious, with the unbalanced phenomenon of the data set itself. Therefore, the two indicators of the model that cannot make gradient adjustments in the prediction are quite different, while the models that can make gradient adjustments, such as GBDT and XGBoost, are relatively balanced.

(3) F1-Score
Considering that the analysis results combining precision and recall rate are not intuitive enough, the F1 score is used to reflect the performance of each model more intuitively. The pairs between models are such as the following Figure 5:

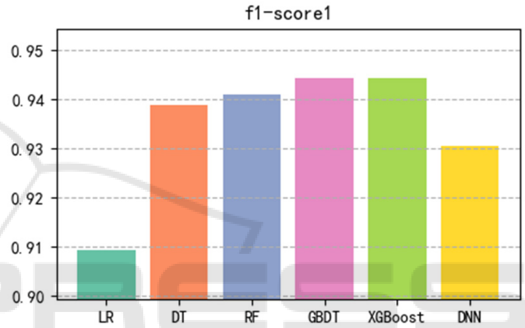

Figure 5: The predicted F1 scores for each model (Photo/Picture credit: Original).

Table 4: The predicted F1 scores for each model.

Method	F1-score
Logistic Regression	0.909
Decision Tree	0.939
Random Forest	0.941
GBDT	0.946
XGBoost	0.95
DNN	0.93

Can intuitively seen from Figure 5, the performance of the GBDT and XGBoost models is better than other models, the reason speculated that there may be two: one because the two models' training mode is more superior, and can constantly adjust data gradient because the experimental data set itself is unbalanced, the two models can more effectively improve the imbalance phenomenon caused by bad results (Table 4). Therefore, the GBDT and XGBoost models have a better performance.

4 CONCLUSION

This study through the UCI data on obesity prediction data research model, then through the model of the data visualization analysis, and by combining the model performance evaluation through the parameters of accuracy, precision, recall rate, and F1 score, finds more accurate obesity prediction model, the results show that GBDT and XGBoost model in the data prediction related indicators are high, the fitting effect is good, the future can through the two machine learning algorithms to predict obesity and related diseases. At the same time can be obtained from the characteristics of age, family whether someone with obesity, whether often eats calorie food and two meals between other food frequencies these four characteristics are associated with obesity, the doctor in determining whether obese patients can be according to these characteristics to further determine whether obesity.

However, there are certain limitations and areas for improvement in this study. The relative lack of data volume in this experiment leads to insufficient analysis of other features. If a large amount of relevant data can be obtained, it will be more favorable for the model prediction.

The experimental data acquisition range is small, and the prediction results are less applicable. The survey results of the data set come from a small range of acquisitions, and it is controversial in its universality. The scope of data collection should be increased to make it have better universality.

REFERENCES

Biau G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13: 1063-1095.

Elreedy D, Atiya A F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences,* 505: 32-64.

Liu Y. The Report on Nutrition and Chronic Diseases of Chinese Residents (2020) was released. *Agricultural Products Market Weekly*, 2021 (2): 58-59.

Liu Y. (2014). Random forest algorithm in big data environment. *Computer modelling & new technologies*, 18(12A): 147-151.

Navada A, Ansari A N, Patil S, et al. (2011). Overview of use of decision tree algorithms in machine learning, *2011 IEEE control and system graduate research colloquium*, 2011: 37-42.

Rigatti S J. (2017). Random forest. *Journal of Insurance Medicine*, 47(1): 31-39.

Sharma S, Sharma S, Athaiya A. (2017). *Activation functions in neural networks*. Towards Data Sci, 6(12): 310-316.

Sharma H, Kumar S. (2016). *A survey on decision tree algorithms of classification in data mining*. International Journal of Science and Research (IJSR), 5(4): 2094-2097.

Song YY, Lu Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry,* 27(2):130-5.

Wang X. (2023). Machine learning based prediction model for heart disease. *Southwestern University*.

Yang D. (2021). Research and application of unbalanced data processing algorithms. *Jiangsu: Jiangsu University of Science and Technology*.

Yu C S, Lin Y J, Lin C H, et al. (2020). Predicting metabolic syndrome with machine learning models using a decision tree algorithm: Retrospective cohort study. *JMIR medical informatics*, 8(3): e17110.

Zhang W, Yu J, Zhao A, et al. (2021). Predictive model of cooling load for ice storage air-conditioning system by using GBDT. *Energy Reports*, 7: 1588-1597.