2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) | 978-1-6654-9246-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/MLISE57402.2022.00099

Comparison Of Different Machine Learning Methods Applied To Obesity Classification

Zhenghao He[1, *]
[1]Department of Computer Science
Tongji University
Shanghai, 200092, China
[*]Corresponding author: 2050259@tongji.edu.cn

***Abstract*—Estimation for obesity levels is always an important topic in medical field since it can provide useful guidance for people that would like to lose weight or keep fit. The article tries to find a model that can predict obesity and provides people with the information of how to avoid overweight. To be more specific, this article applied dimension reduction to the data set to simplify the data and tried to figure out a most decisive feature of obesity through Principal Component Analysis (PCA) based on the data set. The article also used some machine learning methods like Support Vector Machine (SVM), Decision Tree to do prediction of obesity and wanted to find the major reason of obesity. In addition, the article uses Artificial Neural Network (ANN) to do prediction which has more powerful feature extraction ability to do this. Finally, the article found that family history of obesity is the most decisive feature, and it may because of obesity may be greatly affected by genes or the family eating diet may have great influence. And both ANN and Decision tree's accuracy of prediction is higher than 90%.**

Keywords-component; Machine learning; Obesity levels estimation; Dimension reduction

I. INTRODUCTION

Obesity has more than tripled globally since 1975. In 2016, more than 1.9 billion adults aged 18 and older were overweight. More than 650 million of these adults are obese [1]. Obesity has a wide range of health effects, most commonly cardiovascular disease, diabetes, musculoskeletal disease, and some cancers. Many countries have experienced these non-communicable diseases such as obesity and overweight. Although obesity is not a real disease or even sub-health, it has brought deep hidden dangers.

So, it is very essential to predict and prevent obesity. There are many ways to do this. Some researchers do this by using formula, like Body Mass Index (BMI). Some studies did this by using machine learning and used the method like Support Vector Machine (SVM), decision tree, k-means [2, 3]. Some people collected the data and used math formula to predict the obesity rate [4]. In addition, some researchers use three all-cause approaches (partially adjusted, weighted sum, and the two combined) and one cause-of-death approach Comparative Risk Assessment (CRA) to do prediction [5].

According to the [2], the author does prediction by using machine learning. However, the author just used SVM, decision tree, k-means and they did not apply artificial neural network (ANN) that has more powerful feature extraction ability to do this. This paper focused on doing classification, and didn't do dimension reduction to simplify the data, so the most decisive feature of obesity remains unclear. What's more, the research only focused on people aged from 18-25, which is a very small range. According to the [3], the author does estimation by using decision tree and gain a good result. However, the research is only for primary and secondary school students. According to the [4], the author collected data and use formula to predict. However, the data they collected not covered in all aspects.

To solve these limitations mentioned above, this paper had applied Artificial Neural Network (ANN) since it has achieved satisfactory performance in many tasks [6-8], and do dimension reduction through Principal component analysis (PCA), t-SNE, MDS and is suitable for a wider range. The data set used in this paper has 18 variables, which contains more information. And using machine learning can get a more matching model than using formula.

II. METHOD

A. Dataset description and preprocessing

The data comes from this study [9]. It has 2,111 data, 17 features and 7 categories. Table I presents some sample data in the collected dataset.

TABLE I. THE SAMPLE DATA IN THE DATASET.

Features	Data				
	1	*2*	*3*	*4*	*5*
Gender	Female	Female	Male	Male	Male
Age	21	21	23	27	22
Height	1.62	1.52	1.8	1.8	1.78
Weight	64	56	77	87	89.8
family_history_with_overweight	yes	yes	yes	no	no
FAVC	no	no	no	no	no
FCVC	2	3	2	3	2
NCP	3	3	3	3	1
CAEC	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes
SMOKE	no	yes	no	no	no
CH2O	2	3	2	2	2
SCC	no	yes	no	no	no
FAF	0	3	2	2	0
TUE	1	0	1	0	0
CALC	no	Sometimes	Frequently	Frequently	Sometimes
MTRANS	Public_Transportation	Public_Transportation	Public_Transportation	Walking	Public_Transportation
NObeyes	Normal_	Normal_	Normal_	Overwei	Overwe

978-1-6654-9246-1/22/$31.00 ©2022 IEEE
DOI 10.1109/MLISE57402.2022.00099

Features	Data				
	1	*2*	*3*	*4*	*5*
dad	Weight	Weight	Weight	ght_Leve l_I	ight_Le vel_II

Table II is the construction of the feature and the explanation of the abbreviation.

TABLE II. EXPLANATION FOR FEATURES.

Category	Attributes	Explanation
Basic information	Gender	-
	Age	-
	Weight	(kg)
	Height	(m)
	family_history_with_ overweight	Family history with overweight
	SMOKE	Smoke or not
Related to eating habits	FAVC	Frequent consumption of high caloric food
	FCVC	Frequency of consumption of vegetables
	NCP	Number of main meals
	CAEC	Consumption of food between meals
	CH2O	Consumption of water daily
	CALC	Consumption of alcohol
Related with the physical condition	SCC	Calories consumption monitoring
	FAF	Physical activity frequency
	TUE	Time using technology devices
	MTRANS	Transportation used
Dependent variable	NObesity	The level of obesity

This study also considers data processing by replacing and deleting the null value and for easier to process the data, the article changes the string value to int through following way. In the Table III, it shows the integer value corresponding to string type. The left column shows the integer and the column on its right is the value that will convert to it. For example, Male will be converted to 0 when doing data processing. Finally, the article also does normalization to reduce all values in the dataset to between 0 and 1 so that the process of model training can be more stable and faster.

TABLE III. TRANSFORMATION ABOUT FEATURES IN FORMAT OF STRING.

Integer value	Gender	Family history with overweight, FAVC, Smoke, CH2O	CAEC, CALC	MTRANS	Obesity
0	Male	No	No	Public transportation	Insufficient Weigh
1	Female	Yes	Sometimes	Walking	Normal Weight
2			Frequently	Automobile	Overweight Level I
3			always	Motorbike	Overweight Level II
4				Bike	Obesity Type I
5					Obesity Type II
6					Obesity Type III

B. Machine Learning Models

The article did dimension reduction first. First, the article applied PCA to do it. PCA is a method which casts the data to a new coordinate axis which can have the biggest variance of the data, thus finding the most decisive feature of obesity. It is a fast and flexible unsupervised method for dimensionality reduction in data. PCA deals with linear relations.

Then the article used iosmap. Isomap is a non-linear manifold which is based on replacing the Euclidean distance by an approximation of the geodesic distance on the manifold [10]. Isomap can deal with non-linear problem.

In addition, the article used t-SNE to do the task. t-SNE is a method which converts affinities of data points to probabilities. The affinities in the original space are represented by Gaussian joint probabilities and the affinities in the embedded space are represented by t-distributions. t-SNE will focus on the local structure of the data and will tend to extract clustered local groups of samples. It has great effect in reducing the tendency to crowd points together at the center.

When doing PCA, the article set the n_components=0.9 to get the data which remains 90% of information. When doing Isomap, the article checks the reconstruction error to see the best dimension. When doing t-SNE, the article checks the n_components=2 to do the dimension reduction.

The article also used support vector machines (SVM) to do classification. SVM is an algorithm that finds a line to best separation the data. In addition, Gaussian naive bayes was used to classify the data. It is assumed that the continuous values associated with each category are distributed according to Gauss distribution. Decision tree was also used in this article. It can do classification effectively.

C. Deep Learning Models

Artificial Neural network is a widely parallel and interconnected network composed of adaptive simple units. Its organization can simulate the interactive response of biological nervous system to real-world objects. The dropout layer is supposed to preventing the model from overfitting and enhancing the processing speed. The fully connected layer can collect the feature that the former layers processed thus predicting the most possible type of obesity.

After several tests the article enhance the model to do prediction more accurate and the final model's structure can be seen in Figure 1. It has five layers. The Input layer has 16 neurons. The Layer1 has 128 neurons. The Layer2 has 256 neurons. The Layer3 has 32 neurons. The Output Layer has 7 neurons. The Layer2's activation function is sigmoid, the output layer's activation function is Softmax and other layers'

activation function is ReLU. The model's Optimizer is Adam. The evaluation metric is accuracy, and the epoch is 200.

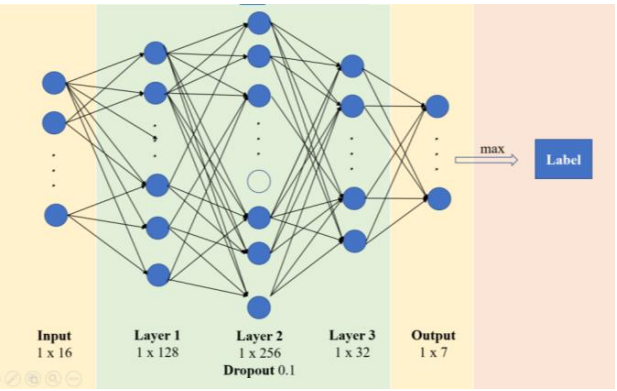


Figure 1. The structure of final ANN model

III. Results And Discussion

A. Dimension Reduction

The article used PCA, Isomap and t-SNE to do the dimension reduction and try to figure out a best model to separate the data set and find the most decisive feature.

1) Principal Component Analysis (PCA)

When doing PCA with the n_component=2, the article visualizes the result, which can be found in Figure 2. And the article also prints the explained variance, which represents the information that is remained after doing PCA. However, the explained variance rate is only 0.1622087 and 0.11519378, which means it cannot remain enough information when using PCA to reduce to 2D. And when doing PCA after changing the code through letting the n_component=0.9, which means remain 90% information of the data set. It shows that the data needs to be reduced to 12 dimensions. When using PCA to reduce the dimension to 1D, the most decisive feature of the data was found through finding the biggest component. The most decisive feature is the family history with obesity. The article provided two reasons to explain this. The first is obesity may pass through genes. The other explanation is people in the same family share the same eating habits, which may result in obesity.

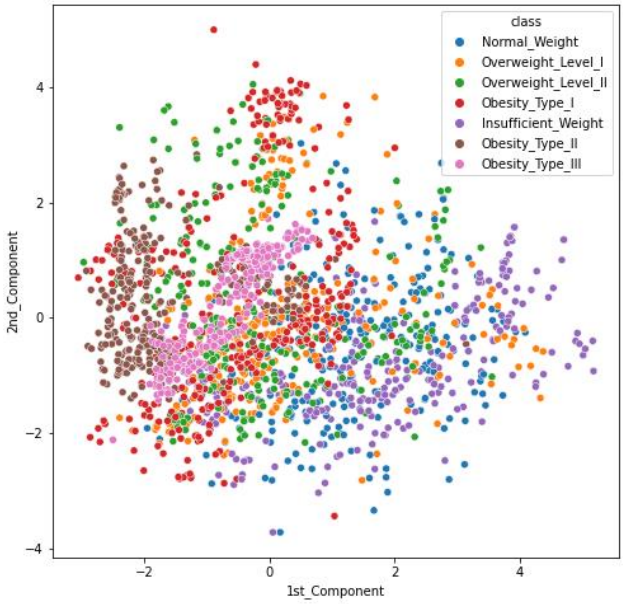
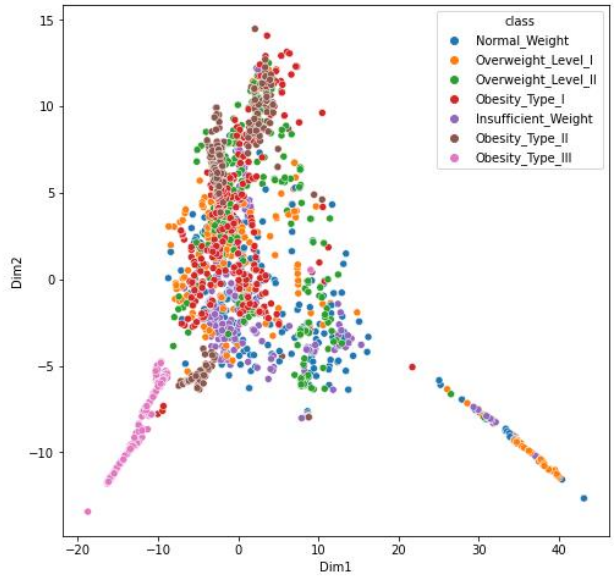


Figure 2. Using PCA to reduce data to 2D

2) Isomap

When using Isomap, the article reduces the data to 2D and 3D and visualize it, which can be found in Figure 3. It can be seen that the result is not very satisfactory, each class mixed together. And the article prints the reconstruction error of each dimension and try to find a dimension that can remain enough information. As can be seen in the Table IV, this study needs to reduce to 12 dimension or more, which is not better than PCA.



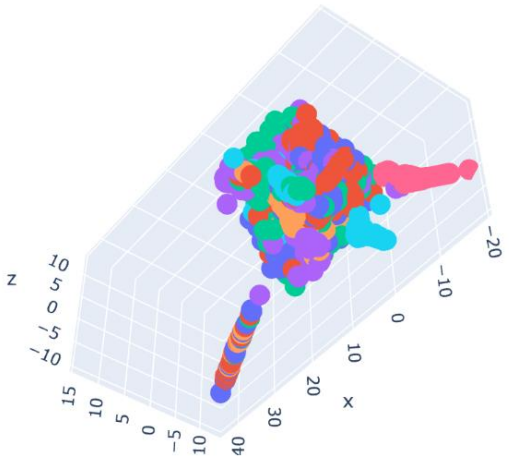


Figure 3. Using Isomap to reduce data to 2D and 3D

TABLE IV. RECONSTRUCTION ERROR RATE OF EACH DIMENSION

dimension	reconstruction error rate
1	58.179
2	44.088
3	36.331
4	31.728
5	28.707
6	26.335
7	25.063
8	24.351
9	23.894
10	23.532
11	23.206
12	22.968
13	22.785
14	22.650
15	22.536

3) t-SNE

When using t-SNE, the article reduces the data to 2D and visualize it. As can be seen in the Figure 4, the classes are separated very well. It is much better than both PCA and Isomap when it needs to separate the class.

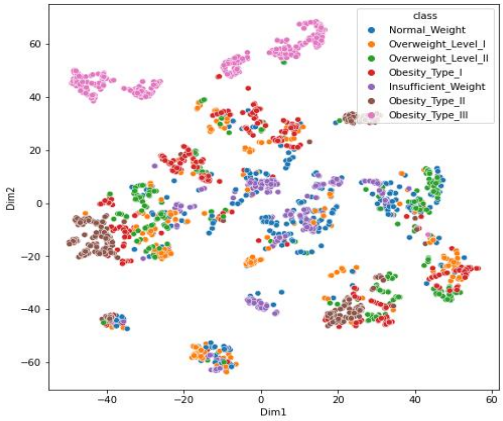


Figure 4. Use t-SNE to reduce data to 2D and 3D

B. Machine Learning

When the article uses SVM to do the classification, the result is shown in Figure 5. While using gaussian naive bayes, the result of training set is shown in Figure 6 and the result of testing set is shown in Figure 7.

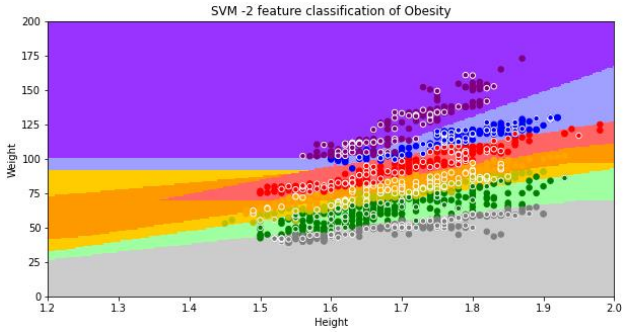


Figure 5. Using SVM to classify data.

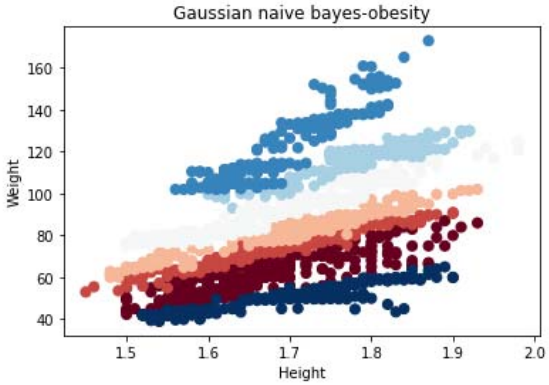


Figure 6. Result of training set when using gaussian naive bayes to classify data.

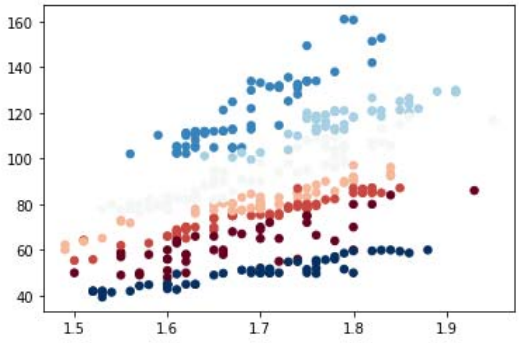


Figure7 Result of testing set when using gaussian naive bayes to classify data.

When using decision tree, the features can be divided into to 3 groups:

1) Attributes related to physical condition: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS).

2) Attributes related to eating habits: Frequent consumption of high caloric food (FAVC), Frequency of consumption of

vegetables (FCVC), Number of main meals, Consumption of food between meals (CAEC), Consumption of water daily (CH20), Consumption of alcohol (CALC).

3) Basic information: gender, age, height, weight and family_history_with_overweight

And the accuracy of each group can be seen in the Table V. The third group has 90% accuracy which is the highest, which means some basic information like family history with overweight have great effect on obesity. This result is similar to the result getting from PCA. May be obesity is hereditary or it is greatly affected by family's eating habit.

TABLE V. ACCURACY FOR EACH GROUP.

Group	Accuracy
1	0.34
2	0.53
3	0.90

C. *Deep Learning*

Neural network is also used in this study to predict dataset. The model used in this article's test accuracy is 0.9195 and its test loss is 0.5182. The Figure 8 is the picture depicts the relationship between loss and accuracy and epoch. When the epoch > 200, the slope of loss and accuracy becomes small. So the final model uses 200 epoch. Figure 9 shows the heatmap of confusion map. It can be observed that most prediction is accurate, which turns out the correctness of the model.

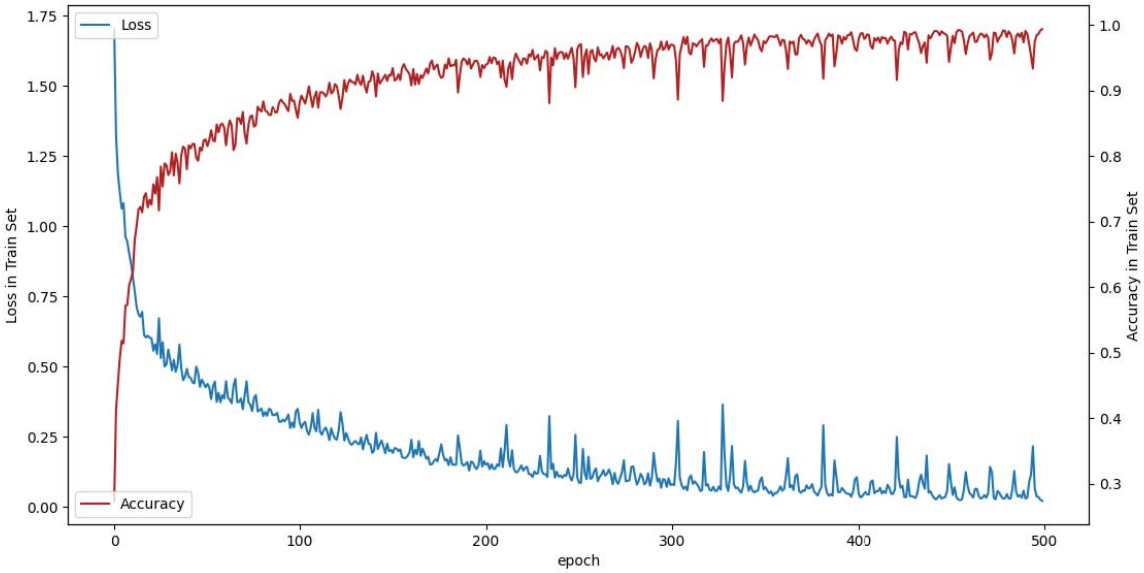


Figure 8. Result of testing set when using gaussian naive bayes to classify data.

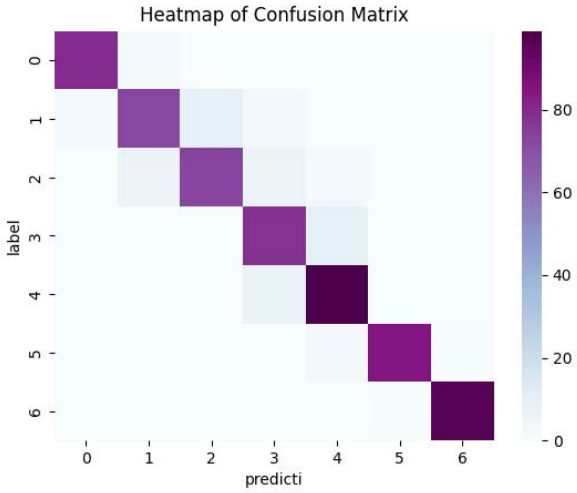


Figure 9. Heatmap of Confusion Matrix

IV. CONCLUSION

In this study, different machine learning algorithms were proposed to estimate the obesity levels. To be more specific, dimension reduction algorithms e.g. PCA, T-SNE and Isomap were employed to present the distribution of data for observation and analysis. Then, machine learning e.g. SVM and deep learning algorithms e.g. ANN was used to predict the obesity. The result shows that proposed scheme can achieve satisfactory performance in this task. In the future, more advanced algorithms will be considered to further improve the accuracy.

REFERENCES

[1] H. Rohana et al. "Worldwide epidemic of obesity." Obesity and obstetrics. Elsevier, pp. 3-8, 2020.

[2] C. Rodolfo Canas, and U. Martinez Palacio. "Estimation of obesity levels based on computational intelligence." Informatics in Medicine Unlocked, vol. 21, 2020.

[3] J. Shi, “Research on prediction model of overweight and obesity of primary and middle school students in Tianjin Based on decision tree method.” Diss. Tianjin Medical University, 2017.

[4] W. Zhao, et al. "An individual level obesity prediction model.", CN102129507A. 2011.

[5] Nikoletta, et al. "Impact of Different Estimation Methods on Obesity-Attributable Mortality Levels and Trends: The Case of The Netherlands. " International journal of environmental research and public health, 2018.

[6] M. Malik and R. Kamra, "A Novel PV based ANN Optimized Converter for off grids Locomotives," 2021 International Conference on Technological Advancements and Innovations (ICTAI), 2021, pp. 299-302.

[7] L. D. Zhang, L. Jia and W. X. Zhu, "Overview of traffic flow hybrid ANN forecasting algorithm study," 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), 2010, pp. V1-615-V1-619.

[8] Y. Qiu, C. S. Chang, J. L. Yan, L. Ko and T. S. Chang, "Semantic Segmentation of Intracranial Hemorrhages in Head CT Scans," 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), 2019, pp. 112-115.

[9] R. Cervantes, and P. U. M. "Estimation of obesity levels based on computational intelligence." Informatics in Medicine Unlocked, vol. 21, 2020.

[10] Y. Bengio, et al. "Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering." Advances in neural information processing systems, vol. 16, 2004.