UTM Computing Proceedings
Innovations in Computing Technology and Applications

Classification of Obesity Level Using Machine Learning

Decision Tree, Random Forest, Support Vector Machine

Ji Tong Lin
School of Computing, Faculty of Engineering
Universiti Teknologi Malaysia (UTM)
Selangor, Malaysia
jitonglin1998@gmail.com

Dr Noor Hidayah binti Zakaria
School of Computing, Faculty of Engineering
Universiti Teknologi Malaysia (UTM)
Johor, Malaysia
noorhidayah.z@utm.my

***Abstract* — Obesity refers to abnormal or excessive fat accumulation in the body. In 2017, it has a global mortality of 4 million deaths. Besides fatality, obese individual suffers from health risks, social-hardships, unemployment, and decreased productivity, which in turn impact negatively on a nation's economy. The study aims to determine machine learning models capable of effective classification of obesity level. To achieve the aim, the study applied 3 different machine learning models, namely decision tree (DT), random forest (RF) and support vector machine (SVM) on the obesity dataset containing the eating habits and physical condition of individuals with several levels of obesity. The result obtained shows that the RF model outperforms the rest of the models with the assessments on accuracy (0.9211), precision (0.9303), recall (0.9211) and F1 score (0.9421). Cross validation using repeated stratified K fold of 10 repeats and 10 splits used to generalize the performance of the models is able to validate this result, in which the RF model showed improved average accuracy recorded at 0.9421.**

Obesity; Machine Learning; Decision Tree; Random Forest; Support Vector Machine (SVM)

I. INTRODUCTION

Overweight and obesity, as defined by [1], are referred as abnormal or excessive accumulation of fat in the body that put the individuals to exposed health risk. Specifically, a body mass index (BMI) of over 25 is labelled overweight, and over 30 is obese. With over 4 million individuals deceased from obesity in 2017, the disease has grown into an epidemic proportion, creating a burden globally.

Individuals with obesity might have their daily lives significantly affected as they are at major risk for developing a range of comorbid conditions, including cardiovascular diseases (CVD), gastrointestinal disorders, type 2 diabetes, just to name a few [2]. Unemployment, social-hardships and increased healthcare costs are some compromises that an obesity individual faces aside from the health risks [3].

It was also found out that obesity leads to impacts on economics of a nation. Individuals with obesity are less productive than individuals with normal weight. Study shows that individuals with obesity is more likely to miss days of work, and work less frequently at full capacity than individuals with normal weight. Moreover, deaths of talented individuals resulted from the worsen health by obesity are likely to negatively affect a nation's economy with the loss of potential contributions by the deceased [4].

Increased physical activity such as exercising and decreased food intake are one of the methods where therapists target on individuals with obesity to effectively relieve the symptoms and eventually serves as a long-term management of the disease. However, such alteration induces a negative energy balance and triggers a cascade of metabolic and neurohormonal adaptive mechanisms, which is challenging to practice [5].

The study realizes that obesity is indeed a worrying health risk with noticeable growth and needs to be stopped. Therefore, the study proposes a solution to the diagnosis of obesity by experimenting with the obesity data containing the individual's eating habits and physical condition with several machine learning models, with the aim of identifying an optimal model capable of the classification of obesity level.

II. RELATED WORKS

The diagnosis of obesity is usually done solely on the BMI or metabolic healthiness. In the research done by [6], machine learning is used whereby unsupervised learning model using clustering was built for the diagnosis of obesity by studying its heterogenous clinical characteristics. From the 4 independent cohorts of clinical data obtained for the study, a mean accuracy of 0.941 was obtained from the classification on the verification set using the developed model.

In [7], the prediction of the BMI of a teenager using their previous BMI values was done using multivariate regression methods and multilayer perceptron (MLP) feed-forward neural network models on the millennium cohort study. An accuracy of over 90% was achieved.

Studies involving several classification models was done in [8]. Prognosis models on the prediction of future risks of diabetes using data involving BMI obtained from Kuwait Health Network (KHN) were built, which involves the use of k-nearest neighbor (KNN), support vector machine (SVM), and logistic regression algorithms. The prediction on the risk of diabetes for 3, 5 and 7 years showed that the KNN model outperformed the rest of the models with AUC values of 0.83, 0.82 and 0.79 for the respective years of prediction.

Some researchers attempted to study obesity without the use of machine learning. In [9], a coding system based on the diagnosis on 4 domains, namely the A (pathophysiology) B (BMI) C (complications remediable by weight loss) and D (degree of severity) is developed. The study was able to address the integral of the 4 mentioned domains as a chronic disease, and proves that scientifically correct and medically actionable approach to disease coding can lead to an effective obesity therapy.

Study on the prevalence of obesity based on country-level demographic, socioeconomic, healthcare and environmental factors was done in [10] using multivariate regression on the data obtained from 3138 counties. According to the result, the variation in obesity prevalence according to the several factors mentioned are 44.9%, 33.0%, 15.5% and 9.1% respectively.

III. METHODOLOGY

The experimental workflow is depicted in Figure 1. The obesity data is first acquired from [11], containing the estimation of obesity level in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The experiment is followed by data pre-processing steps of identifying missing value, categorical data one-hot-encoding, and numerical data normalization. The cleaned Obesity dataset undergoes train-test split with ratio 70:30. In classification, the ZeroR algorithm is first used to identify the baseline performance of the models by using the test dataset. Then, 3 different classification models are built, namely decision tree (DT), random forest (RF), and support vector machine (SVM), to obtain 3 different classification scenarios. Cross validation (CV) is done on all 3 models on the train set to obtain a more generalizable result. This is done by simulating 100 runs using 10 repeated stratified K fold with 10 repeats and 10 splits. Both the classification result on the test set and the CV result are used for model evaluation. Finally, the result obtained are discussed with justifications.

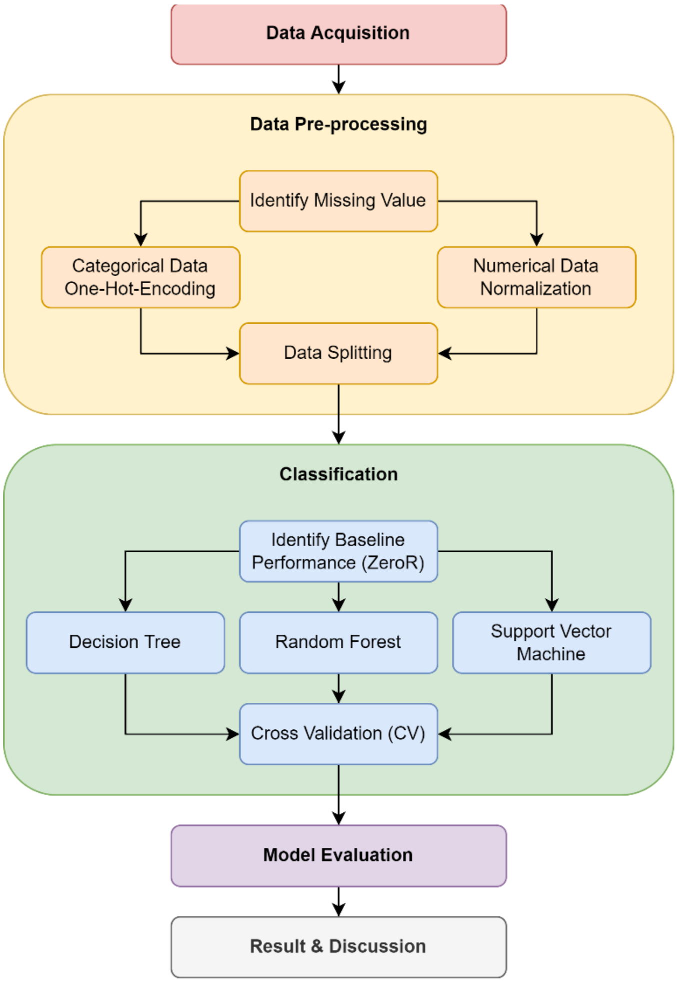


Figure 1. The experimental workflow

A. Data Acquisition

The study starts with the acquisition of the obesity data from [11]. The acquired obesity dataset consists of 2111 samples with 17 features including the target feature "NObeyesdad". According to [11], the obesity dataset is cleaned without any missing values. Besides that, SMOTE was also performed on the dataset by the authors to handle the class imbalance issue.

B. Data Pre-processing

The acquired obesity data undergoes data-preprocessing. Even though the dataset is claimed to be cleaned without missing values, the study still proceeds the pre-processing step with identifying missing values. Indeed, it was found that there are no missing values in the obesity dataset.

The obesity data contains both the numerical data and categorical data, which comprises of 8 features for both type of data, excluding the target feature. Both data types undergo different pre-processing steps. For the categorical data, one-hot-encoding (OHE) is used to encode each unique value in the categorical features into new features with binary values 0 and 1. The original categorical feature is discarded and n number of new features will be introduced if the feature contains n number of unique values. For instance, the "gender" feature has 2 unique values "male" and "female", therefore 2 new features are introduced and the original "gender" feature is discarded. Table I summarizes the number of unique values in each categorical feature (excluding target feature) and the total number of the categorical features after OHE. The obesity dataset now has a total of 31 features after the OHE on the categorical features (including the target feature).

TABLE I. ONE-HOT-ENCODING ON THE CATEGORICAL FEATURES

No.	Feature Name	No. of Unique Value
1	Gender	2
2	family_history_with_overweight	2
3	FAVC	2
4	CAEC	4
5	SMOKE	2
6	SCC	2
7	CALC	4
8	MTRANS	5
Total categorical features after OHE		**23**

The target feature "NObeyesdad" is a multi-class categorical feature with 7 class labels. These class labels are also encoded into an ordinal scale. The result of encoding for the target feature is summarized in Table II.

TABLE II. ENCODING OF THE TARGET FEATURE

No.	Label	Encoded Label
1	Insufficient_Weight	0
2	Normal_Weight	1
3	Obesity_Type_I	2
4	Obesity_Type_II	3
5	Obesity_Type_III	4
6	Overweight_Level_I	5
7	Overweight_Level_II	6
Total class label		**7**

The numerical data on the other hand undergo data normalization. The values are scaled within 0 to 1.

After OHE on the categorical data and normalization on the numerical data, the obesity data is split into train and test set with a ratio of 70:30 with stratification as shown in Figure 2.

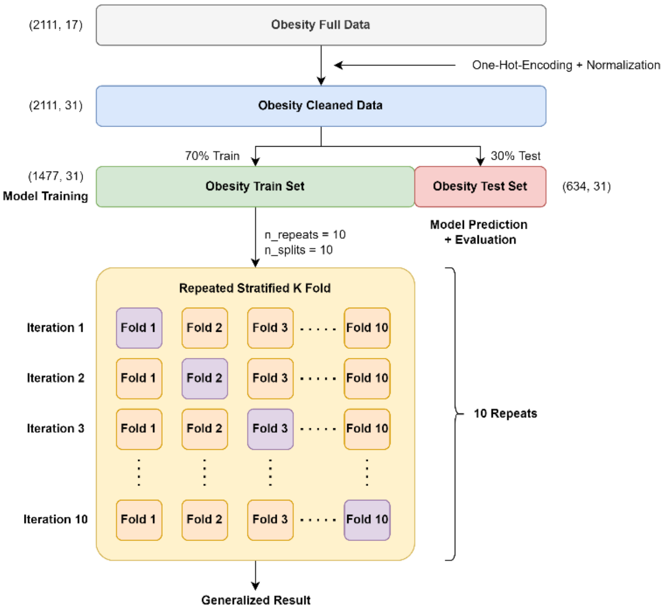


Figure 2. Data Splitting on the Diabetes Dataset

70% of the data will be used for the training of the selected classification models while the remaining 30% will be used to obtain the final performance of the models. The train set is also used for CV to assess the robustness of the models by running the classification 100 times with different folds alternating as train and test set. The class label distribution for both the train and test set is tabulated in Table III.

TABLE III. CLASS DISTRIBUTION OF THE TRAIN AND TEST SET

Class Label	Dataset		
	Train Set	Test Set	Total Samples
0	190	82	272
1	201	86	287
2	245	106	351
3	208	89	297
4	227	97	324
5	203	87	290
6	203	87	290
Total Samples	**1477**	**634**	**2111**

C. *Classification*

The baseline performance of the models is first obtained by using the ZeroR algorithm. The ZeroR algorithm determines the baseline performance of the models by predicting the class label while ignoring the features. By predicting the class label with highest frequency, the baseline accuracy can be obtained via ZeroR.

Next, the study proceeds with the classification of the obesity data using 3 different classification model, namely decision tree (DT), random forest (RF) and support vector machine (SVM). The implementation of the classification models is by using the SciKit Learn (sklearn) package in python. The hyperparameters used for the development of the 3 classification models are kept as default. The classification models are built using the train set, and the final performance of each model is obtained by predicting the class labels on the test set.

D. *Model Evaluation*

To evaluate the model, the test set is used to obtain he classification result from the 3 models. The prediction of the class labels is compared to the actual class labels and the confusion matrix is plotted for each model. From the confusion matrix, several metrics such as the accuracy, precision, recall and F1 score are obtained.

To obtain a more generalizable result, cross validation (CV) is done on the 3 models using the train set. The setting used for the CV is repeated stratified K fold with 10 repeats and 10 folds, resulting in a total of 100 runs.

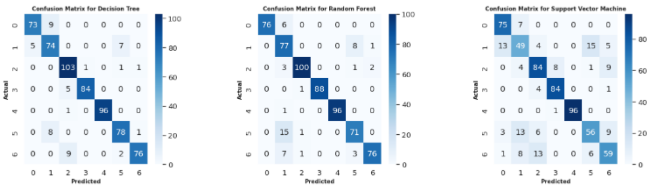
E. *Result & Discussion*

The results obtained from the classification of the obesity dataset using the 3 models are recorded. Discussion on the result is done by comparing the performance of the 3 models using the calculated metrics and the CV result.

IV. Findings

To obtain the highest baseline accuracy, the ZeroR algorithm predicts the majority class label, which is "Obesity_Type_I" or "2" at 106 samples in the test set. The baseline accuracy obtained is 0.1672 or 16.72%.

After training the DT, RF and SVM models using the train set, the prediction of the class label by each model is recorded, and the confusion matrix is plotted in Figure 3.



(a) **(b)** **(c)**

Figure 3. Confusion Matrix for DT (a), RF (b), and SVM (c)

With the confusion matrix, the accuracy, precision, recall and F1 Score are calculated based on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) from the classification. The performance metrics for each model is tabulated in Table IV. Similarly, the repeated stratified K fold with 10 repeats and 10 splits is used for CV and the average accuracy and standard deviation is recorded in Table IV at the bottom row.

TABLE IV. PERFORMANCE METRICS FOR EACH MODEL

Metrics	**Classification Model**		
	DT	**RF**	**SVM**
Accuracy	0.9211[a]	0.9211[a]	0.7934
Precision	0.9244	0.9303[a]	0.7899
Recall	0.9211[a]	0.9211[a]	0.7934
F1 Score	0.9217	0.9236[a]	0.7907
CV Score[b]	0.9266 (0.0218)	0.9431 (0.0200)[a]	0.8313 (0.0267)

a. Superior performance is highlighted in green.

b. CV Score denoted as (average accuracy, standard deviation).

V. Discussion

From the classification result, it can be seen that the DT and RF models perform very similarly, while the SVM model is noticeably behind. Both the DT and RF algorithms achieved the same score for accuracy and recall, which is 0.9211. As for precision and F1 score, the RF model is slightly leading with 0.9303 and 0.9236 respectively compared to the DT model at 0.9244 and 0.9217. With these metrics, the RF model slightly outperforms the DT model. Nevertheless, both the DT and RF model achieved excellent classification result using the obesity dataset. The SVM model on the other hand achieved near 0.8 accuracy, which is a decent result, but is still not as good as the other two models.

To validate the classification result, the CV result for each model is compared better generalize the result. It can be seen that the average accuracy obtained from a total of 100 runs for each model is higher than their respective accuracy on the test set. The difference in accuracy between the classification on test set and CV recorded the highest for the SVM model (0.0379), followed by the RF model (0.0220) and the DT model (0.0055).

Based on the metrics obtained from the classification of each model on the test set and their CV score, it can be concluded that RF model performs the best, followed by the DT model and SVM model.

VI. Conclusion

The study introduced obesity, its variants and its problems or side effects that harms human beings. Several efforts by the other researchers in their study in obesity-related researches are also studied. With that, the study attempts to propose a solution to the classification of obesity level using data consists of the eating habits and physical condition of individuals collected in [11], by using 3 machine learning models, namely decision tree (DT), random forest (RF) and support vector machine (SVM).

After performing appropriate data pre-processing techniques on the obesity dataset, the obesity dataset is split into train and test set. Using the train set, the 3 classification models are trained to learn the features that correlates to the target feature. The test set is then fed into the 3 trained classification models to obtain the performance of each model. By using a confusion matrix, several metrics such as accuracy, precision, recall and F1 score are used to assess the performance of the model in classification. Repeated stratified

K fold with 10 repeats and 10 splits are used to produce a more generalized result by simulating 100 runs of classification using the obesity train set.

Using both the metrics calculated from the confusion matrix and the results obtained from CV, the study concludes that the RF model is the most optimal model suitable for the classification of obesity level using the specific dataset, followed by the DT model and the SVM model.

Acknowledgment

As the author of this study, I would like to thank my lecturer, Dr Noor Hidayah binti Zakaria, for giving me this opportunity to carry out this research. The obesity dataset given to me was supplemented with the research paper associated to it, which has provided me enormous information to work on this research. I would also like to thank my friends who had aided me throughout the research period. Without them, the journey to completing this research would not be as accomplishing.

References

[1] World Health Organization. (n.d.). Health Topic – Obesity. Retrieved from https://www.who.int/health-topics/obesity.

[2] Fruh S. M. (2017). Obesity: Risk factors, complications, and strategies for sustainable long-term weight management. Journal of the American Association of Nurse Practitioners, 29(S1), S3–S14. https://doi.org/10.1002/2327-6924.12510.

[3] Ren, J., Wu, N. N., Wang, S., Sowers, J. R., & Zhang, Y. (2021). Obesity cardiomyopathy: evidence, mechanisms, and therapeutic implications. Physiological reviews, 101(4), 1745–1807. https://doi.org/10.1152/physrev.00030.2020.

[4] Okunogbe, A., Nugent, R., Spencer, G., Ralston, J., & Wilding, J. (2021). Economic impacts of overweight and obesity: current and future estimates for eight countries. BMJ global health, 6(10), e006351. https://doi.org/10.1136/bmjgh-2021-006351.

[5] Wharton, S., Lau, D., Vallis, M., Sharma, A. M., Biertho, L., Campbell-Scherer, D., Adamo, K., Alberga, A., Bell, R., Boulé, N., Boyling, E., Brown, J., Calam, B., Clarke, C., Crowshoe, L., Divalentino, D., Forhan, M., Freedhoff, Y., Gagner, M., Glazer, S., … Wicklum, S. (2020). Obesity in adults: a clinical practice guideline. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne, 192(31), E875–E891. https://doi.org/10.1503/cmaj.191707.

[6] Lin, Z., Feng, W., Liu, Y., Ma, C., Arefan, D., Zhou, D., Cheng, X., Yu, J., Gao, L., Du, L., You, H., Zhu, J., Zhu, D., Wu, S., & Qu, S. (2021). Machine Learning to Identify Metabolic Subtypes of Obesity: A Multi-Center Study. Frontiers in endocrinology, 12, 713592. https://doi.org/10.3389/fendo.2021.713592.

[7] B. Singh and H. Tawfik, "A Machine Learning Approach for Predicting Weight Gain Risks in Young Adults," 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT), 2019, pp. 231-234, doi: 10.1109/DESSERT.2019.8770016.

[8] Farran, B., AlWotayan, R., Alkandari, H., Al-Abdulrazzaq, D., Channanath, A., & Thanaraj, T. A. (2019). Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait. Frontiers in endocrinology, 10, 624. https://doi.org/10.3389/fendo.2019.00624.

[9] Garvey WT, Mechanick JI. Proposal for a Scientifically Correct and Medically Actionable Disease Classification System (ICD) for Obesity. Obesity (Silver Spring). 2020 Mar;28(3):484-492. doi: 10.1002/oby.22727. PMID: 32090513; PMCID: PMC7045990.

[10] Scheinker, D., Valencia, A., & Rodriguez, F. (2019). Identification of Factors Associated With Variation in US County-Level Obesity Prevalence Rates Using Epidemiologic vs Machine Learning Models. JAMA network open, 2(4), e192884. https://doi.org/10.1001/jamanetworkopen.2019.2884.

[11] Palechor, F. M., & Manotas, A. H. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in brief, 25, 104344. https://doi.org/10.1016/j.dib.2019.104344.