Obesity Level Prediction Based on Data Mining Techniques

Asma Alqahtani, ***Fatima* Albuainin**, **Rana Alrayes**, **Noura Al muhanna**, **Eyman Alyahyan**[1] **and Ezaz Aldahasi**[2]

Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, P.O. Box 31961, Jubail, Kingdom of Saudi Arabia

Summary

Obesity affects individuals of all gender and ages worldwide; consequently, several studies have performed great works to define factors causing it. This study develops an effective method to trace obesity levels based on supervised data mining techniques such as Random Forest and Multi-Layer Perception (MLP), so as to tackle this universal epidemic. Notably, the dataset was from countries like Mexico, Peru, and Colombia in the 14- 61year age group, with varying eating habits and physical conditions. The data includes 2111 instances and 17 attributes labelled using NObesity, which facilitates categorization of data using Overweight Levels l I and II, Insufficient Weight, Normal Weight, as well as Obesity Type I to III. This study found that the highest accuracy was achieved by Random Forest algorithm in comparison to the MLP algorithm, with an overall classification rate of 96.7%.

Key words:

Obesity, Data Mining, prediction, Multilayer Perceptron (MLP), Random Forest.

1. Introduction

The obesity epidemic is universally prevalent [1]. It contributes to exacerbate many chronic ailments like kidney disease, cardiovascular, and cancer, etc. [2] [3]. Obesity refers to an excessive accumulation of body fat in certain areas of the body that can be harmful to health. It is found in adults, teenagers, and children [4]. Biological risk factors, such as genetic history, are known to cause obesity. Other risk factors such as psychological and social habits and eating can also not be ruled out. Since 1980, the number of obese individuals globally has doubled, and in 2014, more than 1900 million adults experienced a change in their weight. Some of the causes of weight gain are increased intake of energy-dense, high-fat foods and decreased physical activity caused by sedentary work, new transportation modes, and increasing urbanization [5].

Despite many attempts to reduce obesity through exercise/diet, raising awareness, surgery, and drug therapy, an effective solution has not yet been found to and diagnose it accurately at an early stage. Data mining has been utilized in information technology for medical decision-making, such as prognostic and diagnostic problems, and for detecting correlations between the risk factors and outcomes [6]. Data Mining (DM) analyzes many data to discover unknown patterns and extract hidden information [7]. DM can be categorized into descriptive and predictive tasks. The descriptive task focuses more on describing the data, grouping it into categories, and summarizing it. On the other hand, the predictive task analyzes historical data and produces patterns/conclusions for future predictions [8].

Machine learning techniques (ML) find numerous applications in the domain of DM. Many studies have predicted and analyzed obesity using web tools[9][10][11][12][13][14][15]. However, these studies are typically confined to calculating BMI and omitting as associated factors like family background and the time allocated to it. To the best of our knowledge, only one study forecasts obesity that considers family background [16]. They primarily depend on three modeling methods: decision trees (J48), naïve bayes, and logistic regression.

This study aims to predict obesity using Random forest and multi-layer perceptron (MLP). It also aims to determine the most important predictive factors with a significant effect on obesity. The main contribution of this study is to carry out a comparative analysis of previous algorithms via a recent dataset used in [17]. The results will go a long way in addressing the obesity issue and enhancing the health prospects of individuals.

This paper is structured in the following manner. Section 2 elaborates on this work's literature Section 3, describes proposed techniques. Section 4 focuses on Empirical studies, whereas Section 5 shows the optimization strategy. Section 6 presents results and discussion. Section 7 further expounds discussion highlights. Finally, Section 8 provides the conclusion and future works.

2. Related Work

In [16], the problem of obesity was considered, and the disease was analyzed before producing web tools. The SEMMA data mining methodology was undertaken to pick, model, and explore the dataset. Next, three techniques, Bayesian networks, Logistic Regression, and Decision trees, were chosen. The decision trees were found to have the best outcome on the basis of metrics: precision, TP rate, FP rate, and recall. Using WEKA, the Decision Trees technique was observed to have the best precision rate of 97.4%.

In [9], the authors considered obesity that has been growing steadily in children, teenagers, and adults. In this study, a

Manuscript received March 5, 2021
Manuscript revised March 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.3.14

computational intelligence-based method was put forth, utilizing Decision Trees, Supportive Vector Machines (SVM) and K-Means – data mining techniques. The dataset was selected from 81 male students and 97 female students in the (18 - 25) age group from Colombia, Peru, and Mexico. A comparative examination was done to improve the proposed tool. Then, the best approach was combined with the clustering technique after obtaining the results of classification techniques.

In [10], the authors took into consideration the importance of defining people at risk of being affected by obesity as quickly as possible and quickly/accurately predict possible BMI rates for young adults from data pertaining to early childhood in the Millennium Cohort Study (MCS) using machine learning techniques. Various experiments were done using multivariate regression algorithms as well as multi-layer perceptron feed-forward artificial neural networks (MLPFFANN). The MLPFFANN technique obtained better results than regression algorithms, with over 90% prediction accuracy.

In [11], the authors utilized data mining techniques for forecasting the obesity's risk factors in Bangladesh through middle ages. The study proposed the risk mining technique (PRMT) to forecast obesity class-based risk factors. The class level precision, evaluation method and the data analysis results rely on WEKA's software using different machine learning algorithms. The study collected data from rural and urban regions regarding different risk factors concerning daily activities. The Naïve Bayes technique was found to yield the best results via 10-fold cross-validation.

In [12], the authors explored the challenges of predicting health diseases. They proposed improved machine learning models to forecast obesity: binary logistic regression, improved decision tree IDT, ،weighted k-nearest neighbor KNN and artificial neural network ANN. When comparing the binary logistic regression model with an accuracy of 56.02%, KNN, IDT, and ANN were found to be significantly better, as the accuracy of the KNN model was 88.82%. The IDT model had an accuracy of 80.23%, whereas it was 84.22% for the ANN model.

In [13], the authors examined body fat percentage (BFP) and the high cost of its measuring devices. This study aimed to define the BFP using hybrid machine learning classifiers at a high rate and minimum parameters. To that end, they generated four various hybrid models with SVM, MLFFNN, and DT. This study's practical outcome showed that the produced system could be utilized to predict the BFP. The system is also capable of measuring BFP with a single anthropometric measurement.

In [14],the current traditional methods' high cost to gather data on public health was considered with a view to predicting obesity. This study used the data mining tool of WEKA data and carried out predictive analytics via the J48 classifier so as to ascertain the rate of accuracy. The result confirmed the J48 classifier had predicted obesity from patterns of calorie consumption with a precision of 89.41%.

In [15], this study adopted a country-level approach to gauge obesity's spreading rate using domestic sales of food and beverages categories (a subset). The study introduced three machine learning algorithms for non-linear regression. This study ascertained data and obesity prevalence for 79 nations. The proposed method was validated with regard to both proportions of the countries and the absolute prediction error where a satisfactory obesity prevalence was forecasted. It was observed that baked goods and flour, accompanied by carbonated beverages and cheese, are the most significant food type to forecast g obesity.

Data mining is known to be the pivotal factor to curb this global disease which damages the health of people across all age groups. Therefore, a review of previous literature found that many sources share the same goal of detecting and reducing obesity levels using data mining techniques. However, they differed in terms of databases and the target age range. From Table 1, it can be seen that similar algorithms were used to deal with different databases such as Decision Trees, Supportive Vector Machines (SVM), K-Mean, along with other algorithms. Therefore, we used random forest and multi-layer perception algorithms in our study to achieve the best accuracy rates to trace levels of obesity.

Table 1: Previous studies highlight obesity disease.

Ref	***Year***	***Proposed Method***	***Dataset***	***Best Method***
[16]	2019	J48 NB LR	Attributes: 18 Instances: 712 Related to: The undergraduates ages of (18 - 25) from (Colombia, Mexico, and Peru).	DT precision rate of 97.4 %
[9]	2020	DT SVM K-Means	Related to: (18 - 25) years, 81 males and 97 females from institutions in Colombia, Peru, and Mexico.	DT ROC Area 98.2 %
[10]	2019	Multivariate linear regression Linear SVM, Quadratic SVM Fine Tree Ensemble Bagged Trees Trees	Related to: Young Adults (14 years old and older). The data used: From the Millennium Cohort Study (MCS).	MLPFFANN prediction accuracy rate of 93.4%

		MLPFFANN		
[11]	2018	NB IBK KStar ZeroR Random tree Simple logistic	Related to: Middle-Aged People from Bangladesh. The data used: From specified urban and rural areas.	NB accuracy rate of 99.2 %
[12]	2017	KNN IDT ANN BLR	Related to: high school students (in grades 9-12) The data used: High schools in Tennessee.	ANN accuracy rate of 99.46%
[13]	2020	MLFFNN SVM DT	It is related to body fat percentage values and anthropometric measurements of 252 individuals.	MAPE Performance MLFFNN + DT + SVMs
[14]	2018	J48	Related to: Malaysian grocery data, demographic data, and anthropometric data.	J48 accuracy rate of 89.4118%
[15]	2019	SVM Random Forest Extreme gradient boosting.	Related to: beverage and food sales data in 48 categories for 79 nations.	Baked goods, flour, carbonated drinks as well as cheese, are most pertinent-food category to forecast obesity.

3. Description of Proposed Techniques

3.1 Artificial Neural Network

An artificial neural network (ANN) involves performance attributes that are similar to the biological neural networks of human brains [18].ANN is capable of identifying non-linear linkages between the data set's inputs and outputs. As modeling tools, it is practical and useful, particularly in difficult problems to explain through statistical and physical equations [18]. In this context, one of the neural networks that is used most extensively is Multilayer Perceptron (MLP) which primarily comprises nodes -artificial neurons- organized into three types of layers, namely, input nodes, hidden nodes, and output nodes. Fig. 1 illustrates the network process of an MLP [20].

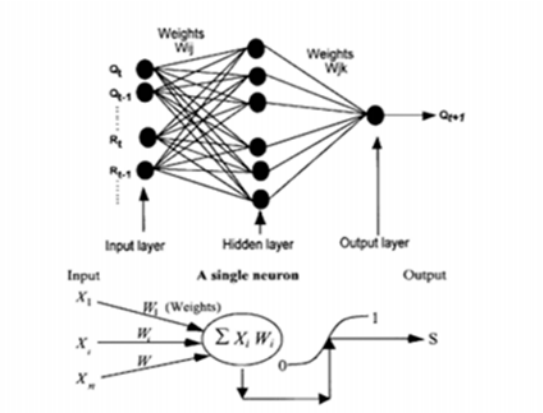

Fig 1 Illustration of an MLP network [20]

3.2 Random Forest

The random forest model consists of numerous individual decision trees working as an ensemble. Leo Breiman introduced this algorithm in 2001 on the basis of a regression tree [21]. The random forest approach is broadly used in classification problems [22][23] due to its power and ability to manage extensive features with small samples. As shown in Fig.2, the bootstrap sample method is applied to train each tree of the training data set. This technique seeks a random subset of variables to split in each node. For classification, the input vector of each unit is fed into the RF and each tree votes for a class. The RF finally selects the target with maximum votes. It can manage enormous input data sets unlike other models[24].

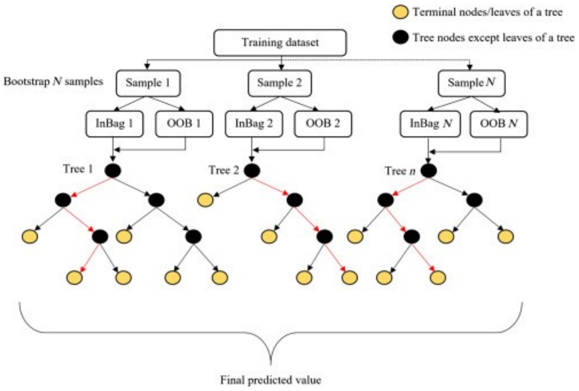

Fig 2 Structure of a random forest [24]

4. Empirical Studies

4.1 Description of the Dataset

The dataset includes data to predict obesity of who depend on their diet and lifestyles belonging to Mexico, Colombia, and Peru. We used an internet platform to collect data (23%) from users; we then used Waikato Environment for Knowledge Analysis (WEKA) tool to create 77% of the data. The data consist of 17 attributes (see Table 2) and 2111 records. The approach of data preparation is summarized in Fig. 3. NObesity, the class variable makes it possible to classify data (using values Normal Weight, Insufficient Weight, Obesity Type I/Type II/Type III, and Overweight Level I/ Overweight Level II

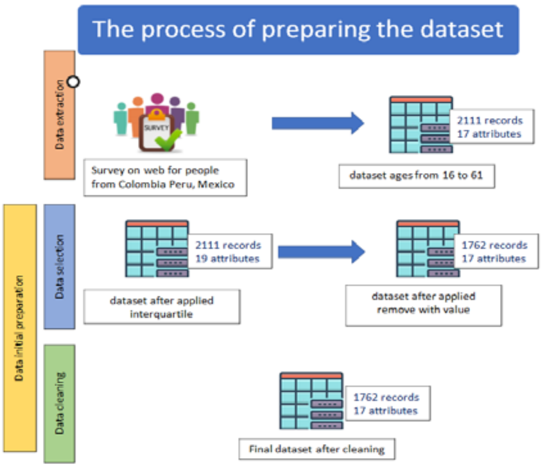
The process of preparing the dataset
Data extraction
SURVEY
Survey on web for people from Colombia Peru, Mexico
2111 records 17 attributes
dataset ages from 16 to 61
Data initial preparation
Data selection
2111 records 19 attributes
dataset after applied interquartile
1762 records 17 attributes
dataset after applied remove with value
Data cleaning
1762 records 17 attributes
Final dataset after cleaning

Fig 3 Preparing dataset process.

Table 2 Attribute's description

Attribute	***Description***	***Initial value***
Gender	The gender of obesity-prone condition	Male Female
Age	The age of the obesity-prone case	Integer Numeric Values
Height	The duration of the obesity-prone case	Integer Numeric Values (Mt)
Weight	The weight of the obese condition	Integer Numeric Values (Kg)
Family history with Obesity	an obesity-prone family or no	Yes No
FAVC	Consumption of calorie-rich food on a frequent basis	Yes No
FCVC	Usage of vegetables (frequency)	Always Sometimes Rarely
NCP	Number of meals	to 2 1 3 More than 3
CAEC	Consumption of food between meals	Always Usually Sometimes Rarely
CH20	Consumption of water daily	Less than one liter More than 2 liters Between 1 and 2 liters
CALC	Consumption of alcohol	Less than one liter More than 2 liters Between 1 and 2 liters
SCC	Calorie's consumption monitoring	Yes No
FAF	Physical activity frequency	to 2 days 1 to 4 days 3 to 6 days 5 No physical activity
TUE	Time using technology devices.	0to 2 hours 3 to 5 hours
MTRANS	Transportation used.	MTRANS Transportation used Public transportation Motorbike Bike Walking Automobile
Smoking	Determining if obesity-prone individual smokes or not	Yes No
NObesity	The class variable facilitates the data classification so as to classify a person's obesity level and recommend systems monitoring obesity levels.	Insufficient Weight, Normal Weight Obesity Type I, Obesity Type II Obesity Type II Overweight Level I, Overweight Level II

4.2 Experimental Setup

Obesity prediction was made using WEKA, machine learning software and toolkit. The Java framework was distributed under the GNU General Public License. The tool presents several modern and popular techniques for analyzing and mining data. WEKA is able to support many data mining activities to forecast health problems, such as data preprocessing, classification, grouping, simulation, correlation, and functional choice. [25]. The data contains many issues that need to be preprocessed, such as missing values, outliers, irrelevant or redundant data, etc. This is one of the most important data mining phases that helps clean the data to be used as input to the other processes [26].

To begin with, the database is prepared and preprocessed for the experiment. Then, the outliers and extreme values in the dataset were detected using WEKA's unsupervised attribute filter (Interquartile range). After determining instances with outliers or extreme values, we got rid of these instances from dataset through the RemoveWithValues filter. It was necessary to increase the models' accuracy by overcoming outliers' problem.

Next, the optimization parameters for Random Forest and MLP were determined to obtain better performance as well as to guarantee optimal results. This step entailed adjusting the Seed, Numerations factors for Random Forest, while Seed and Hidden layers and the Learning Rate for MLP. As shown in Table 3 and Table 4, the performance was better with the default value of the parameters. Random Forest had an accuracy of 96.70%, while the MLP had an accuracy of 95.06%.

Moreover, we determine the relationship of information gain and coefficients between class variables and calculated all attributes so as to get the characteristics of selected features ranked. Tables 5 and 6 show the findings.

Subsequently, to predict if one person suffers from obesity, the capability to increase the precision of classification performance was examined by ascertaining the important features. Feature selection/correlation-based features selection were studied based on Info Gain. The results are shown in Table 7 and Table 8.

Following the gathering of results, the classifiers were implemented using numerous ratios of partition. The Random Forest and MLP attained a ratio accuracy of 70:30 (70 % for data training as well as 30 % for testing). Findings are shown in Table 9.

Eventually, based on the above experimental results, we generated the final model by selecting the most effective subset features achieving the optimal cross-validation or partition ratio.

5. Optimization Strategy

In order to enhance the classification results and to obtain accuracy-based better performance, the Weka meta-learner (CV Parameter Selection) search methodology was used [27]. Table 3 shows the default and optimum parameters for each classifier, and it was found that the optimum parameters for the classifiers are the default parameters. Table 4 depicts the classifiers' performance using optimal parameters.

Table 3 Default and optimal parameters for each classifier-Values

Model	***Parameters Values***		
	Parameters	***Default value***	***Optimal value***
Random Forest	numIterations	100	100
	Seed	1	1
MLP	Seed	0	0
	Hidden Layers	a	a
	Learning Rate	0.3	0.3

Table 4 Default and optimal parameters for each classifier-Performance

Model	***Performance (accuracy)***
	Default value (Optimal value)
Random Forest	96.70%
MLP	95.06%

6. Results and Discussion

6.1 Feature Selection Impact on Dataset

The Information Gain (InfoGain) [28], and correlation-based features selection method [29],were used to select the best performing subset, besides the most significant attributes with the highest impact to forecast obesity.

As Table 5 shows, the coefficient's correlation was taken into consideration for ranking characteristics on the basis of Pearson values with the involvement of class variable (output). In addition, Table 6 shows that InfoGain was implemented to classify the features on the basis of measure, of information gain with the involvement of class variable (output).

Table 7 and Table 8 present the results of the Info Gain and correlation-based features selection method. The best performance was observing to require the use of each features. Furthermore, accuracy was found to reduce upon reducing the number of factors.

Table 5. Each attribute and the class- Correlation

Order	***Attribute's name***	***Correlation with class***
1	Weight	0.342
2	Family_history_with_overweight	0.1997
3	Gender	0.1955
4	CAEC	0.1734
5	FCVC	0.1725
6	CALC	0.1552
7	Age	0.1501
8	FAVC	0.1287
9	NCP	0.1253
10	Height	0.1069
11	MTRANS	0.0982
12	SCC	0.0921
13	FAF	0.0832
14	CH2O	0.0682
15	TUE	0.0645
16	SMOKE	0.0558

Table 6. Each attribute and the class – Information gain

Order	***Attribute's name***	***Infogain with class***
1	Weight	1.7517
2	Age	0.8016
3	FCVC	0.6259
4	FAF	0.4839
5	TUE	0.4346
6	CH2O	0.4153
7	Gender	0.3471
8	NCP	0.344
9	Height	0.2563
10	Family_history_with_overweight	0.2163
11	CAEC	0.2058
12	CALC	0.169
13	MTRANS	0.1172
14	FAVC	0.0881
15	SCC	0.0437
16	SMOKE	0.015

Table 7. Correlation-based feature selection results

Number of features	***Features***	***Random Forest***	***MLP***	***AVG***
All features	*All*	96.70 %	95.06 %	95.88 %
Nine *features*	*Weight* Family_history_with_overweight Gender CAEC FCVC CALC Age FAVC *NObesity*	91.25 %	83.71 %	87.48 %
Five features	*Weight* Family_history_with_overweight Gender CAEC *NObesity*	77.46 %	76.2 %	76.83 %
Four features	*Weight* Family_history_with_overweight Gender *NObesity*	76.16 %	74.97 %	75.56 %
Three features	*Weight* Family_history_with_overweight *NObesity*	67.8%	61.5 %	64.65 %

Table 8. Infogain feature selection results

Number of features	***Features***	***Random Forest***	***MLP***	***AVG***
All features	*All*	96.70%	95.06%	95.88%
Nine *features*	Weight Age FCVC FAF TUE CH2O Gender NCP *NObesity*	92.67%	79.17%	85.92%
Five features	Weight Age FCVC FAF *NObesity*	88.76%	72.6%	80.68%
Four features	Weight Age FCVC *NObesity*	87.06%	69.46%	78.26%
Three features	Weight Age *NObesity*	83.65%	64.2%	73.9%

6.2 Effect of Different Partition Ratios on The Dataset

Following the identification of optimal features, it was evident that each feature assumed significance in InfoGain as well as the correlation-based features selection method. The performance of each classifier was evaluated by performing many experiments on the data using various ratios of partition in the 50-80 range. Table 9 illustrates the findings of each classifier's direction participations.

Table 9. Results of different partition ratios

Partition ratio	***Performance***	
	Random Forest	***MLP***
50:50	94.55%	91.6%
60:40	95.03%	91.48%
70:30	94.51%	93.38%
80:20	94.6%	93.18%

6.3 Direct Partition Techniques and 10-fold validation– Comparison

During the comparison of the above two methods, it was observed that the (10-fold validation method helped obtain an improved value as compared to the direct partition ratio. As shown in Table 10.

Table 10. Direct partition techniques versus 10-fold cross validation

Techniques	***Proposed model***	
	Random Forest	***MLP***
10-fold validation	96.70%	95.06%
Partition ratio	94.51%	93.38%

7. Further Discussion

Table 11 shows that the ultimate model to forecast obesity was developed using each feature via optimum parameters. By using the method of 10-fold cross-validation, we obtained optimal results for each Random Forest and MLP. Random Forest was found to outperform MLP to forecast obesity disease with a satisfactory accuracy of 96.70%. The classification performance could also be enhanced by using each feature (16 of them) by making use of the optimal criteria for each classifier, as illustrated in Fig. 4. Through this process, we identified attributes that significantly affected the ability to predict Type II obesity disease: Weight Age, FCVC, FAF, TUE, CH2O, Gender, NCP, Height, Family_history_with_overweight, CAEC, CALC, MTRANS, FAVC, SCC, and SMOKE.

Another indicator of the performance of the classification model is the Receiver Operating Characteristic (ROC) curve. The placement and proximity of this curve to the left-hand side (on the top) reveals the high accuracy level of this experiment. Overall, the area under the curve for each classifier shown in Fig. 5 and Fig. 6 shows that the most suitable classifier is determined to be Random Forest in comparison to MLP.

The proposed method was found to surpass the findings achieved by [10] with 93.4% precision, whereas [14] showed a precision of 89.41 %.

The results can investigate the importance of computational intelligence-based approaches to research various diseases or pathologies, detecting them early and adequately, while minimizing their societal effects

Table 11. Performance of proposed model.

Techniques	***Proposed Model***	
	Random Forest	***MLP***
10-fold validation	96.70%	95.06%

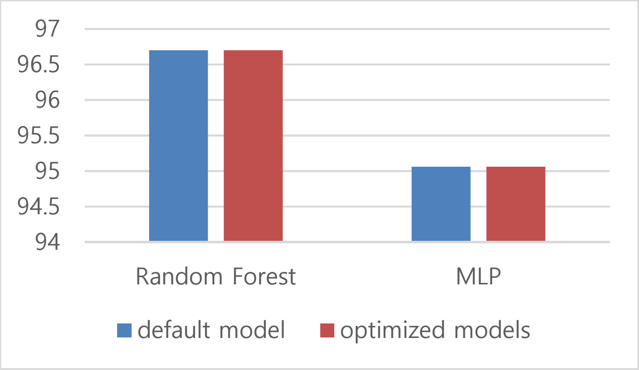

Fig 4. Default and optimized models – Comparison

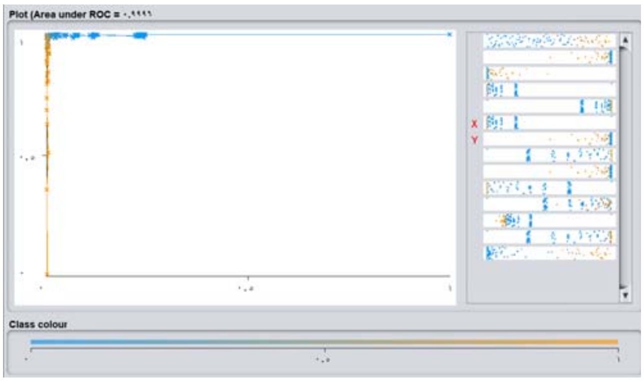

Fig 5. Random Forest curve - Obesity Type II class

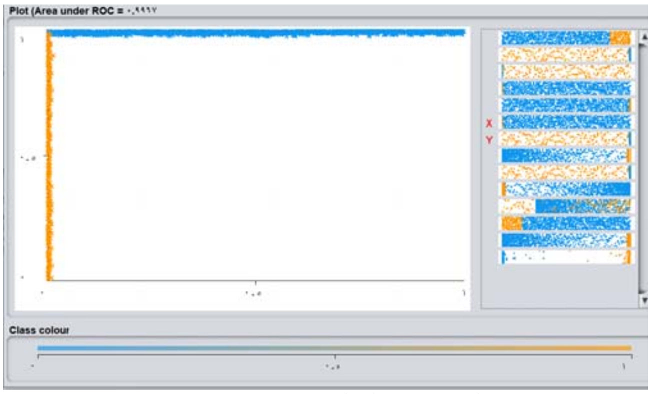

Fig 6. MLP ROC curve - Obesity Type II class

8. CONCLUSION

Data mining involves carrying out data analysis with a view to distinguishing knowledge behaviors or patterns. Researchers have used this discipline to develop solutions for addressing numerous societal issues, for example disease identification on the basis of historical data. Analysis of obesity assumes importance as it is a global menace damaging the lives of millions of people worldwide, regardless of gender or age. Various authors have devoted time and effort to examine this issue and define the pathology providing a theme relating to continuous evolution. This study applied two techniques, random forest and MLP, to obtain the best accuracy rates to predict obesity. Consequently, the random forest method outperforms MLP to predict obesity at an early stage with high accuracy of 96.70%. The experiment showed that the algorithms work better when using all features; accordingly, all 16 features were used to achieve the highest accuracy ratio. The findings were found to surpass those in earlier studies [10] [14] with precision values of 93.4% and 89.41 %, respectively. This study's results will facilitate the evaluation of the importance of computational intelligence-based approaches to accurately and correctly examine various diseases or pathologies, diagnose them on a timely basis, and lower the adverse impact of such diseases on society.

In the future, it could be possible to expand this research to other feature selection approaches with a view to enhancing the obesity production's accuracy and evaluation, also potentially replicating or adopting other machine learning algorithms. The dataset could also possibly be expanded in the future. Making use of massive data can be beneficial in the health sector. Quite a few studies have been carried out in this field in Saudi Arabia. A comprehensive data collection to develop Saudi-based models will make a significant contribution to Saudi development.

References

[1] A. Bewick and E. P. Greener, "Ref 1.Pdf." p. 4623, 1969.

[2] H. B. Hubert, M. Feinleib, P. M. McNamara, and W. P. Castelli, "Obesity as an independent risk factor for cardiovascular disease: A 26-year follow-up of participants in the Framingham Heart Study," *Circulation*, vol. 67, no. 5, pp. 968–977, 1983, doi: 10.1161/01.CIR.67.5.968.

[3] A. Must, J. Spadano, E. H. Coakley, A. E. Field, G. Colditz, and W. H. Dietz, "The disease burden associated with overweight and obesity," *J. Am. Med. Assoc.*, vol. 282, no. 16, pp. 1523–1529, 1999, doi: 10.1001/jama.282.16.1523.

[4] B. Guy-Grand, "Beyond body mass index," *Cah. Nutr. Diet.*, vol. 49, no. 3, pp. 93–94, 2014, doi: 10.1016/j.cnd.2014.05.002.

[5] E. Alyahyan and D. Dusteaor, "Decision trees for very early prediction of student's achievement," *2020 2nd Int. Conf. Comput. Inf. Sci. ICCIS 2020*, 2020, doi: 10.1109/ICCIS49240.2020.9257646.

[6] N. Lavrač, "Selected techniques for data mining in medicine," *Artif. Intell. Med.*, vol. 16, no. 1, pp. 3–23, 1999, doi: 10.1016/S0933-3657(98)00062-1.

[7] M. H. J. and P. Jian and Kamber, "Data Mining Techniques, Third Edition," p. 847, 2011.

[8] M. Khajehei and F. Etemady, "Data mining and medical research studies," *Proc. - 2nd Int. Conf. Comput. Intell. Model. Simulation, CIMSim 2010*, no. September 2010, pp. 119–122, 2010, doi: 10.1109/CIMSiM.2010.24.

[9] R. C. Cervantes and U. M. Palacio, "Estimation of obesity levels based on computational intelligence," *Informatics Med. Unlocked*, vol. 21, no. November, 2020, doi: 10.1016/j.imu.2020.100472.

[10] B. Singh and H. Tawfik, "A Machine Learning Approach for Predicting Weight Gain Risks in Young Adults," *Conf. Proc. 2019 10th Int. Conf. Dependable Syst. Serv. Technol. DESSERT 2019*, pp. 231–234, 2019, doi: 10.1109/DESSERT.2019.8770016.

[11] R. Hossain, S. M. H. Mahmud, M. A. Hossin, S. R. Haider Noori, and H. Jahan, "PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques," *Procedia Comput. Sci.*, vol. 132, pp. 1068–1076, 2018, doi: 10.1016/j.procs.2018.05.022.

[12] Z. Zheng and K. Ruggiero, "Using machine learning to predict obesity in high school students," *Proc. - 2017 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2017*, vol. 2017-Janua, pp. 2132–2138, 2017, doi: 10.1109/BIBM.2017.8217988.

[13] M. K. Uçar, Z. Uçar, F. Köksal, and N. Daldal, "Estimation of body fat percentage using hybrid machine learning algorithms," *Meas. J. Int. Meas. Confed.*, vol. 167, 2021, doi: 10.1016/j.measurement.2020.108173.

[14] N. Daud, N. L. Mohd Noor, S. A. Aljunid, N. Noordin, and N. I. M. Fahmi Teng, "Predictive Analytics: The Application of J48 Algorithm on Grocery Data to Predict Obesity," *2018 IEEE Conf. Big Data Anal. ICBDA 2018*, pp. 1–6, 2019, doi: 10.1109/ICBDAA.2018.8629623.

[15] J. Dunstan, M. Aguirre, M. Bastías, C. Nau, T. A. Glass, and F. Tobar, "Predicting nationwide obesity from food

sales using machine learning," *Health Informatics J.*, vol. 26, no. 1, pp. 652–663, 2020, doi: 10.1177/1460458219845959.

[16] E. De-La-Hoz-Correa, F. E. Mendoza-Palechor, A. De-La-Hoz-Manotas, R. C. Morales-Ortega, and S. H. B. Adriana, "Obesity level estimation software based on decision trees," *J. Comput. Sci.*, vol. 15, no. 1, pp. 67–77, 2019, doi: 10.3844/jcssp.2019.67.77.

[17] F. M. Palechor and A. de la H. Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico," *Data Br.*, vol. 25, p. 104344, 2019, doi: 10.1016/j.dib.2019.104344.

[18] A. S. Nur, "Artificial Neural Network Weight Optimization: A Review," *Telkomnika*, 2014.

[19] R. A. Flauzino, *Artificial Neural Networks A Practical Course*. Springer International Publishing, 2016.

[20] S. Riad, J. Mania, L. Bouchaou, and Y. Najjar, "Predicting catchment flow in a semi-arid region via an artificial neural network technique," *Hydrol. Process.*, vol. 18, no. 13, pp. 2387–2393, 2004, doi: 10.1002/hyp.1469.

[21] Y. Qi, "Random forest for bioinformatics," Springer, 2012, pp. 307–323.

[22] P. M. Chakraborty Sounak, Khalilia Mohammed, "Predicting disease risks from highly imbalanced data using random forest," vol. 11, no. 1, p. 51, 2011.

[23] A. F. in A. N. L. Sarica, Alessia; Cerasa, Antonio; Quattrone, "Random Forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review," 2017.

[24] and S. H.] A. Hemmati-Sarapardeh, A. Larestani, M. Nait Amar, *Chapter 2 - Intelligent models*. 2020.

[25] B. J. Saleh, A. Y. F. Saedi, A. T. Q. Al-aqbi, and L. A. Salman, "A Review Paper: Analysis of Weka Data Mining Techniques for Heart Disease Prediction System," *Libr. Philos. Pract.*, vol. 7, no. 1, p. 1, 2020.

[26] R. Sangeetha and S. Sathappan, "Preprocessing Using Attribute Selection in Data Stream Mining," *Proc. 3rd Int. Conf. Commun. Electron. Syst. ICCES 2018*, no. Icces, pp. 431–438, 2018, doi: 10.1109/CESYS.2018.8723918.

[27] R. Kohavi, "Wrappers for performance enhancement and obvious decision graphs," no. September, 1995.

[28] C. M. Lai, W. C. Yeh, and C. Y. Chang, "Gene selection using information gain and improved simplified swarm optimization," *Neurocomputing*, vol. 218, no. November 2018, pp. 331–338, 2016, doi: 10.1016/j.neucom.2016.08.089.

[29] M. Mursalin, Y. Zhang, Y. Chen, and N. V. Chawla, "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier," *Neurocomputing*, vol. 241, no. February, pp. 204–214, 2017, doi: 10.1016/j.neucom.2017.02.053.