NTU Journal of Agricultural and Veterinary Sciences (2023) 3 (4) : 183-191
DOI: https://doi.org/10.56286/ntujavs.v2i2

IRAQI Academic Scientific Journals

JAVS NTU

P-ISSN: 2788-9890 E-ISSN: 2788-9904
NTU Journal of Agricultural and Veterinary Sciences
Available online at: https://journals.ntu.edu.iq/index.php/NTU-JAVS/index

Decision tree in Obesity Level Classification of Northern Technical University Students,

1st Sahar Esmaiel Mahmood [1] ,2nd Ibrahim K. Sarhan [2], 3rd Hadi .M . Ehymmayed [3]
1.Computer Department, Northern Technical University, Mosul, Iraq.
2. Medical Laboratory Department, Alnoor College University, Mosul, Iraq.
3. Medical Laboratory Department, Alnoor College University, Mosul, Iraq.

Article Informations

Received: 14-10- 2023,
Accepted: 04-11-2023,
Published online: 28-12-2023

Corresponding author:
Name: Ibrahim K. Sarhan
Affiliation : Alnoor College University
Email:
ibrahim.khalil@alnoor.edu.iq

Key Words:
Obesity,
Classification,
Data Mining,
Dissection Tree,
Roc Curve

ABSTRACT

The rising prevalence of overweight and obesity among youth and adolescents is a concerning trend in many countries. This poses a significant threat to current and future healthcare systems due to the associated risks of cardiovascular disease, type 2 diabetes, metabolic disorders, and even mortality. Developing effective strategies for preventing these conditions and understanding their origins are crucial. Creating predictive models for overweight and obesity in young individuals and their related outcomes holds immense value, and machine learning models have proven to be valuable tools for this purpose. The main objective of this study is to construct a data-driven model that can forecast the likelihood of overweight or obesity in youngsters. HTo achieve this, the researchers employed Decision tree analysis using the Rapid Miner program. This analysis aimed to determine the extent to which various human variables contribute to obesity and how accurately these factors can classify individuals into different obesity index categories based on their determined body mass (body weight in kg divided by the square of body length in meters). For this study, a body mass of <= 24.9 kg.m-2 was considered normal weight (obesity index = 0), while a body mass above 24.9 kg.m-2 indicated overweight (obesity index = 1), The attributes related to student activity and nutrition were utilized as inputs for the Decision tree models, and the outputs were the obesity index classifications. The results of the investigation demonstrated successful classification of obesity levels, with an efficiency rate of 94.16%. This indicates that the data attributes used in the study were highly accurate in determining the obesity index.

Introduction

Predicting the risks of health conditions and events is a crucial objective in medical and public health research. Healthcare providers utilize these risk factors to identify patients at higher risk within a broader population. Based on evidence of causality, these risk factors can also serve as therapeutic targets. For instance, elevated total cholesterol levels have been identified as a risk factor for cardiovascular disease [1]. Consequently, patients are often prescribed cholesterol-lowering medications to reduce their cholesterol levels. Similarly, diets rich in saturated fat have been associated with obesity [2], leading to recommendations to reduce dietary fat intake for weight loss [3].Obesity is a growing global health problem, with more than 30% of adults classified as obese [4]. Obesity is linked to various diseases, including type 2 diabetes and heart disease [5]. Much of the current research on obesity focuses on the energy balance perspective [6], which results from consuming more calories than the body requires for metabolic energy. On the caloric side, specific dietary components have been identified as risk factors for weight gain. Particularly, dietary fat has been linked to higher body fat and obesity [7, 8], especially saturated fats [9].Recent studies have shown that low-carb diets can be just as effective as low-fat diets in reducing body fat [10].

Traditional epidemiological analysis techniques

such as linear and logistic regressions, have been highly successful in identifying risk factors and causes of specific diseases, especially when a single risk factor has a strong isolated association with a clinical outcome (e.g., evidence linking cigarette smoking to lung and other cancers). However, these approaches have been less effective in explaining complex health outcomes, particularly multifaceted "lifestyle diseases" like cardiovascular disease, metabolic disorder, and obesity. This limitation is partly due to the methodological constraints of these conventional approaches.

Regression models can include terms that account for interactions between two or more factors, as the interplay of multiple factors can be closely related to health research. For example, there is a widespread belief that consuming high-mercury fish poses significant neurological risks [12], especially for pregnant women [13]. However, recent research indicates that the selenium content in fish can counteract this risk [14]. In other words, the interaction or combination of mercury and selenium is more critical than individual factors in determining the neurological risk associated with fish consumption.

Approach

The utilization of machine learning methods offers a valuable complement to traditional regression-based analyses. Machine learning encompasses a diverse range of techniques which can be broadly categorized based on whether the learning is supervised (using outcome data during training). A key characteristic of these methods is their data-driven and experimental nature allowing them to transcend the limitations of linear information disclosure. One significant advantage of machine learning is its ability to identify intricate patterns within the data without explicitly specifying them beforehand. Unlike traditional regression analyses, which may require prior identification of specific relationships, machine learning approaches have an exploratory aspect. They can uncover relationships and patterns in the data that might have been overlooked in linear or logistic regression analyses. Machine learning models excel at recognizing complex patterns making them more effective in predictive tasks or classification tasks. These patterns contain richer information than the sum of individual independent variables alone. As a result, pattern-based classification techniques often outperform traditional regression-based methods in terms of predictive accuracy and classification performance Regenerate responce.

Works utilized

This study presents evidence supporting the use of an approach to predict obesity based on comprehensive data regarding diet, exercise, and pharmacological factors. The data were collected through a survey administered to a group of students from the Northern Technical University - medical departments. Various factors, such as diet nature, fast food consumption, intake of fatty foods, exercise habits, smoking, alcohol and soda consumption, cortisone use, social interactions, and extra workload, were considered in this study. Additionally, physical assessments were conducted. These data were utilized as inputs for a Decision tree analysis, with the obesity index derived from body mass serving as the output. The objective was to classify and predict the specific factors contributing to obesity considering both dietary choices and other human activities. The study aimed to assess the effectiveness of these factors in classifying different obesity levels within the university student community and to provide recommendations for weight management and risk reduction.collected through a survey administered to a group of students from the Northern Technical University - medical departments. Various factors, such as diet nature, fast food consumption, intake of fatty foods, exercise habits, smoking, alcohol and soda consumption, cortisone use, social interactions, and extra workload, were considered in this study. Additionally, physical assessments were conducted. These data were utilized as inputs for a Decision tree analysis, with the obesity index derived from body mass serving as the output. The objective was to classify and predict the specific factors contributing to obesity, considering both dietary choices and other human activities. The study aimed to assess the effectiveness of these factors in classifying different obesity levels within the university student community and to provide recommendations for weight management and risk reduction.

Material and Techniques

This study involved data collected from young college students between the ages of 18 and 25 years attending Northern Technical University. The sample size comprised 137 records, including 79 males and 58 females who participated in the survey. To collect data, a series of questions were administered to identify their obesity levels, considering factors such as age, weight, gender, frequency of physical activity, fast food consumption, and others that could help understand the patterns among overweight individuals. All measurements were taken during the same session to minimize variations in environmental conditions Weight and height measurements were taken with participants barefoot and in light clothing. Body Mass Index (BMI) was calculated using the formula BMI =

The data collected from the survey underwent various types of analysis to identify patterns related to factors influencing obesity in young students. Techniques such as Decision tree models were used to classify and categorize the levels of obesity based on BMI data, as shown in Table 1.

Dataset The main factors contributing to the development of overweight were found to be high Calorie intake, reduced energy expenditure due to a lack of physical activity, genetic factors, economic elements, and potential stress or depression. These factors were identified through a dataset containing 19 variables gathered by the authors through surveys applied to undergraduate students at Northern Technical University in Mosul, Iraq. Table 2 presents the variables considered in determining the levels of overweight and their corresponding.

Table 1. BMI classification according to WHO And Mexican normativity (DO, 2010) [18].
BMI Classification.

Underweight	-1	Less than 18.5
Normal	0	18.5 to 24.9
Overweight	1	25.0 to 29.9
Obesity I	2	30.0 to 34.9
Obesity II	3	35.0 to 39.9
Obesity III	4	Higher than 40

Table 2. Data description

Variable	Detail
Age	Integer Numeric
Gender	1: Male 0: Female
Sports exercise	1: Continues 0:Never
Deal with computer	1: Yes 0: No
Smoking use	1: Yes 0: No
The nature of sleep	1: Normal 0: Not normal
Fast food	1: Yes 0: No
Eat in order	1: Yes 0: No
Eat fast	1: Fast 0: slow
Alcohol intake	1: Yes 0: No
Drink soft drinks	1: Yes 0: Never
Family Genetic History/Obesity	1: Yes 0: No
Gland deficiency	1: Yes 0: No
Resorting to food	1: Yes 0: No
Social relations	1: Yes 0: No
Take cortisone medications	1: Yes 0: No
Eating fatty foods	1: Yes 0: No
Lifestyle	1: Normal 0: Not normal
Working out	1: Yes 0: No
Waist measurement	numeric value cm
Weight	numeric value kg
Height	numeric value cm
Body mass (BMI)	numeric value $kg.m^{-2}$
Obesity index	1:obesity 0: Normal

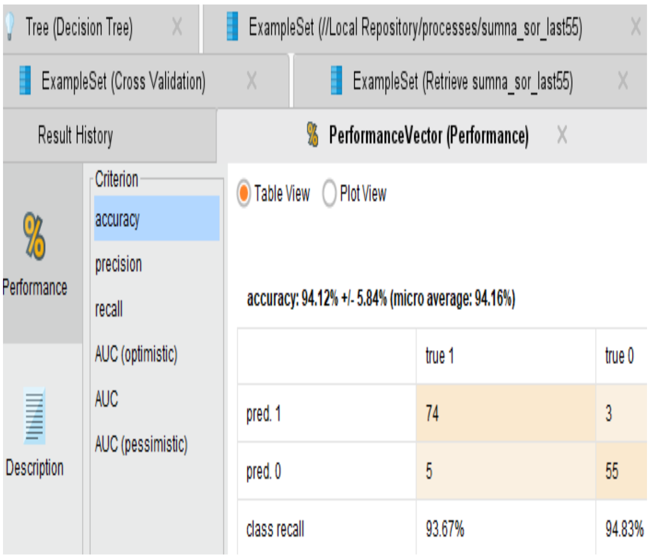
Result and Conversation

In the realm of classifier models, specifically Decision Trees, Edwards et al. [19] propose utilizing a confusion matrix to evaluate the accuracy of obesity level classification. Table 3 represents the confusion matrix which plays a vital role in determining the percentage values of key metrics such as Accuracy, Precision, Sensitivity, and Specificity. Equations 1, 2, 3, and 4, involve True Positive (TP) the actual value was positive and the model predicted a positive value, True Negative (TN) the actual value was negative and the model predicted a negative value, False Positive (FP) the prediction is positive and it is false. (Also known as the Type 1 error), and False Negative (FN) prediction is negative, and result it is also false. (Also known as the Type 2 error), they are used to calculate these metrics. Based on precision, TP rate, FP rate, and recall, the decision trees exhibit the most favorable performance [21].

Table3. The four outcomes can be formulated in a 2×2 Confusion *Matrix*

		Predicted condition	
	Total population	Predicted Condition positive	Predicted Condition negative
True condition	*condition positive*	*True positive TP*	*False positive FP Type I error*
	condition negative	False negative FN Type II error	True negative TN

Table 4. Result of Confusion matrix for Tree Dissection Analysis



	true 1	true 0
pred. 1	74	3
pred. 0	5	55
class recall	93.67%	94.83%

Table 5. Result of Equation, of Accuracy, Precision, Sensitivity, Specifity for Decision Tree Analysis

Accuracy	Precision	Sensitivity	Specifity
94.16%	93.936%	96.10%	91.66%

Equations, formulas

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots\ldots (1)$$

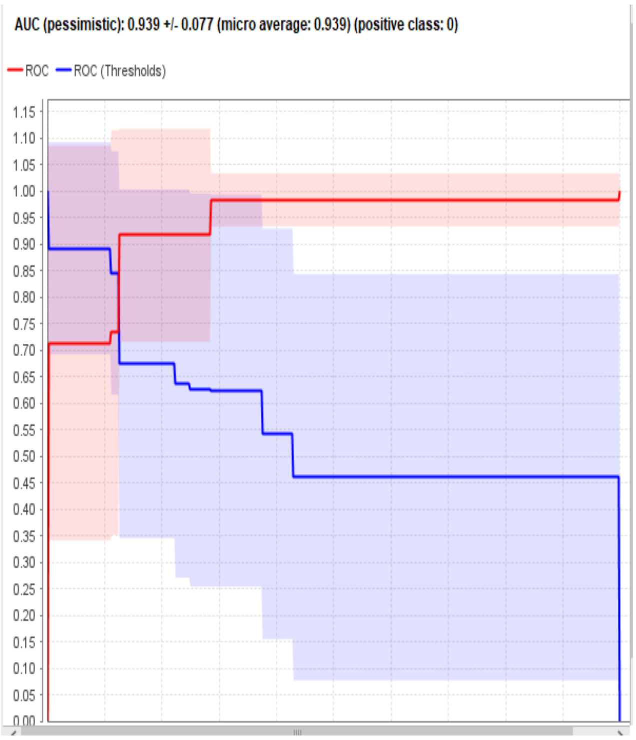
$$Precision = \frac{TP}{TP+FP} \ldots\ldots\ldots\ldots\ldots (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \ldots\ldots\ldots\ldots (3)$$

$$Specifity = \frac{TN}{TN+FP} \ldots\ldots\ldots\ldots\ldots (4)$$

Table 4. displays Result of Confusion matrix for Decision Tree Analysis revealing a notable Accuracy rate of 94.16%. This rate is computed using Equation (1) with data from the Confusion Matrix. Meanwhile, Table 5 showcases the metrics for Accuracy, Precision, Sensitivity, and Specificity, which are observed to be 94.16%, 93.67%, 96.10%, and 91.66%, respectively.

Figure 1 depicts the ROC curve resulting from the Decision Tree Analysis, featuring an AUC value of 0.939 This AUC value is computed using Equations (5) and (6) with data from the Confusion Matrix. The ROC curve showcases how effectively the model can differentiate between different classes at various threshold settings. A higher AUC value indicates the model's proficiency in distinguishing class 0 from class 1, demonstrating its effectiveness in identifying individuals with overweight and those who are normal . The outcomes suggest that the Decision Tree Analysis model performs exceptionally well in classifying normal weight and overweight individuals based on nutrition and human activity factors, as studied by Martínez et al. (2009). Furthermore, the ROC curve in Figure 1 aligns with the data from Confusion Matrix Table 4, serving as an essential performance measurement for the machine learning classification problem with multiple classes. Table 4 provides a 2×2 Confusion Matrix, containing predicted and actual value combinations, which proves highly valuable in assessing sensitivity, precision, specificity, accuracy, and notably, the AUC-ROC curve.



Figuer 1. Roc curve for Dissection Tree Analysis

True Positive Rate (TPR) $= \frac{\Sigma\, True\ positive}{\Sigma\, Condition\ Positive!} = \frac{TP}{TP+FN}$ ... (5)

False Positive Rate (FPR) $= \frac{\Sigma\, False\ positive}{\Sigma\, Condition\ Negative} = \frac{FP}{FP+TN}$ .. (6)

It is noted from Figure 2. related to the dissection tree analysis of the data of students of the Northern Technical University of Mosul for the purpose of classification of the obesity index for university students, whether it is normal (0) or abnormal (1) obesity depending on human activity, nutrition, lifestyle and other daily activities such as smoking, exercise, soft drinks, the nature of food and the nature of social relations, etc. Where it is noted that the most important variable that controls the strong impact of the obesity index is the lifestyle (Life Style), where it is located at the first level in the decision tree, and it is either a normal quiet lifestyle (>0.50) or an abnormal lifestyle in which the person is tense and under stressful pressure (<0.50). Followed in the second level by all of the variables exercise sports activity as well as thyroid disorders In the third level, comes both alcoholic beverages and smoking, followed by the fourth level of importance of the abuse of curtzion drugs, while in the fifth level comes the intake of soft drinks and comes in the sixth level both eating fatty foods , and always return to food. They are followed in later levels by sports activity, the nature of sleep and social relations. All these variables, according to their levels, contribute to determining the obesity index among students (normal, obese).

depending on human activity and dietary habits in life in order to avoid many of the pathological problems that accompany obesity, for example, heart disease, diabetes and blood pressure, in addition to many problems associated with the state of obesity and obesity for members of any society. Therefore, giving recommendations to avoid many activities and dietary habits that lead to raising the obesity factor of the body is of great importance to avoid many diseases and problems that lead to human society damage and reduce its efficiency and production.It can save the disbursement amounts of money spent to treat obesity in hospitals and health resorts. It is noted from the figure for the analysis of the decision tree that lifestyle is the most important variable in distinguishing the obesity factor, as it appears that the decrease in its value to less than (0.50) indicates that the lifestyle is abnormal as (normal: 1 abnormal: 0) and then the indicator moves to the right side of the variable of exercise, where if its value is less than (0.50), the indicator goes to obesity directly, and this indicates that the pattern of tricks is tense with not practicing any sports activity that leads to obesity.

As for practicing sports activity while ensuring that you do not smoke, the trend leads to a normal state, thus avoiding the obesity factor despite the fact that the lifestyle is abnormal (anxiety, tension, psychological fatigue), and thus the great importance of avoiding smoking and practicing sports activity appears. On the left side of the decision tree, despite the fact that the lifestyle is normal and calm, any of the following factors (thyroid disorders, drinking alcohol, taking cortizione drugs) may leads directly to obesity, but in the absence of a disorder of the thyroid gland and the lack of alcohol and curtision intake, the influential factor to raise obesity is both the abuse of soft drinks and the trend of fatty foods. This leads to obesity, but in the case of avoiding fatty foods despite the abuse of soft drinks, the trend is not to be exposed to obesity. It is also noted that in the case of lack of direction for soft drinks, the factors affecting the trend to obesity are (return to food always, lack of exercise, lack of sleep) and the last factor for the establishment of social relations accompanying the lack of exercise and abnormal sleep, this leads to obesity as well. Accordingly, the importance of the decision tree in dealing with data, analyzing it and deriving the right decisions has been approved.

In contrast, exercising can help prevent significant weight gain leading to obesity. These insights highlight the importance of the decision tree in analyzing data and making informed decisions

In Figure 3.the ideal scenario is depicted where two curves do not overlap at all, indicating a perfect level of distinguishability for the model. In this situation, the model is highly capable of accurately identifying both the positive class and the negative class.

As observed in Figure 4, when two distributions overlap, it results in type 1 and type 2 errors. The occurrence of these errors depends on the threshold set for classification, and there is the flexibility to minimize or increase them based on the chosen hreshold. When the AUC (Area Under the Curve) is 0.7, it indicates a 70% probability that the model will successfully distinguish between the positive class and the negative class.

In Figure5, the representation shows the worst-case scenario. When the AUC is approximately 0.5, it indicates that the model lacks discrimination ability to distinguish between the positive class and the negative class. In other words, the model's predictive power is minimal in this situation, and it struggles to differentiate between the two classes effectively.

Conclusion:

The findings of this study revealed significant risk factors associated with weight gain prioritized by their significance through Decision Tree Analysis. These factors include lifestyle choices, gland deficiencies or irregular thyroid secretions, soft drink consumption, intake of fatty foods, age, and lack of exercise. Considering the importance of these variables for overweight individuals is essential.

The Decision Tree classifier models showcased exceptional performance in terms of Accuracy, Precision, and Specificity for classifying individuals into weight categories (normal or overweight). Moreover, the area under the curve (AUC) of the ROC curve reached a value of 0.939, further confirming the high classification ability of the Decision Tree model.

Recommendations for further studies, it would be beneficial to explore additional feature selection methods to enhance the accuracy and evaluation of obesity prediction. This may involve replicating or adopting other machine learning algorithms. Furthermore, expanding the dataset could yield more insights and improve the effectiveness of obesity prediction. Emphasizing the utilization of massive data in the health sector holds great potential for advancing research and better understanding obesity-related issues. By harnessing the power of extensive datasets, significant strides in combating obesity and promoting overall well-being can be achieved successfully.

References

[1] Goff DJ, Lloyd-Jones D, Bennett G, Coady S, D'Agostino RBS, Gibbons R, Greenland P, Lackland D, Levy D, O'Donnell CRJ, Schwartz J, Smith SJ, Sorlie P, Shero S, Stone N, WIlson P (2014) 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a Report of the American College of Cardiology/American Heart Association Task Force on practice guidelines. Circulation 129(suppl 2): S49–S73.

https://doi.org/10.1161/01.cir.00004377414860 6.98

[2] Crescenzo R, Bianco F, Mazzoli A, Giacco A, Cancelliere R, di Fabio G, Zarrelli A, Liverini G, Iossa S 2015 Fat quality influences the obesogenic effect of high fat diets. Nutrients 7(11):9475–9491. https://doi.org/10.3390/nu7115480

[3] U.S. Department of Health and Human Services, U.S. Department of Agriculture (2015) 2015–2020 dietary guidelines for Americans, 8th edn.

[4] Ogden CL, Carroll MD, Fryar CD, Flegal KM 2015 Prevalence of obesity among adults and youth: United States, 2011–2014. NCHS data brief, vol 219, Hyattsville, MD

[5] National Institutes of Health 1998 Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults. vol NIH Publication No. 98-4083. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, and National Heart, Lung, and Blood Institute.

[6] Hill JO, Wyatt HR, Peters JC 2012 Energy balance and obesity. Circulation 126(1):126–132. https://doi.org/10.1161/circulationaha.111.087213

[7] Satia-Abouta J, Patterson RE, Schiller RN, Kristal AR 2002 Energy from fat is associated with obesity in U.S. men: results from the prostate cancer prevention Trial. Prev Med 34 (5):493–501. https://doi.org/10.1006/pmed.2002.1018

[8] Tucker LA, Kano MJ 1992 Dietary fat and body fat: a multivariate study of 205 adult females. Am J Clin Nutr 56(4):616–622

[9] Crescenzo R, Bianco F, Mazzoli A, Giacco A, Cancelliere R, di Fabio G, Zarrelli A, Liverini G, Iossa S 2015 Fat quality influences the obesogenic effect of high fat diets. Nutrients 7(11):9475–9491. https://doi.org/10.3390/nu7115480

[10] Walker TB, Parker MJ 2014 Lessons from the war on dietary fat. J Am Coll Nutr 33(4):347–351. https://doi.org/10.1080/07315724.2013.870055

[11] Satia-Abouta J, Patterson RE, Schiller RN, Kristal AR 2002 Energy from fat is associated with obesity in U.S. men: results from the prostate cancer prevention Trial. Prev Med 34 (5):493–501. https://doi.org/10.1006/pmed.2002.1018

[12] Carocci A, Rovito N, Sinicropi MS, Genchi G 2014 Mercury toxicity and neurodegenerative effects. Rev Environ Contam Toxicol 229:1–18.

https://doi.org/10.1007/978-3-319-03777-6_1

[13] Solan TD, Lindow SW 2014 Mercury exposure in pregnancy: a review. J Perinat Med 42 (6):725–729. https://doi.org/10.1515/jpm-2013-0349

[14] Ralston NV, Ralston CR, Raymond LJ 2016 Selenium health benefit values: updated criteria for mercury risk assessments. Biol Trace Elem Res 171(2):262–269. https://doi.org/10.1007/ s12011-015-0516-z

[15] Cristianini N, Shawe-Taylor J 2000 An introduction to support vector machines and other Kernel-based learning methods. Cambridge University Press, Cambridge, UK

[16] Harrell F 2015 Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. In: Springer series in statistics. Springer

[17] Gómez, M. and L. Ávila, 2008. La obesidad: un factor de riesgo cardiometabólico. Medicina de Familia.

[18] DO, 2010. NORMA Official Mexicana NOM-008-SSA3-2010, Para el tratamiento integral del sobrepesoy la obesidad. Diario Official.

[19] Edwards, W. and B. Fasolo, 2001. Decision technology. Annual Rev. Psychol., 52: 581-606.

[20] Martínez, R.E.B., N.C. Ramírez, H.G.A. Mesa, I.R. Suárez and M.D.C.G. Trejo et al., 2009. Árboles de decisión como herramienta en el diagnóstico médico. Revista Médica de la Univ. Veracruzana, 9: 19.24.

[21] E. De-La-Hoz-Correa, F. E. Mendoza-Palechor, A. DeLa-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.

[22] Z Ge, Z Song, SX Ding, B Huang - Ieee Access, 2017 - ieeexplore.ieee.org

D. Data Mining and Analytics Now, suppose the machine learning procedure is completed, and the data model

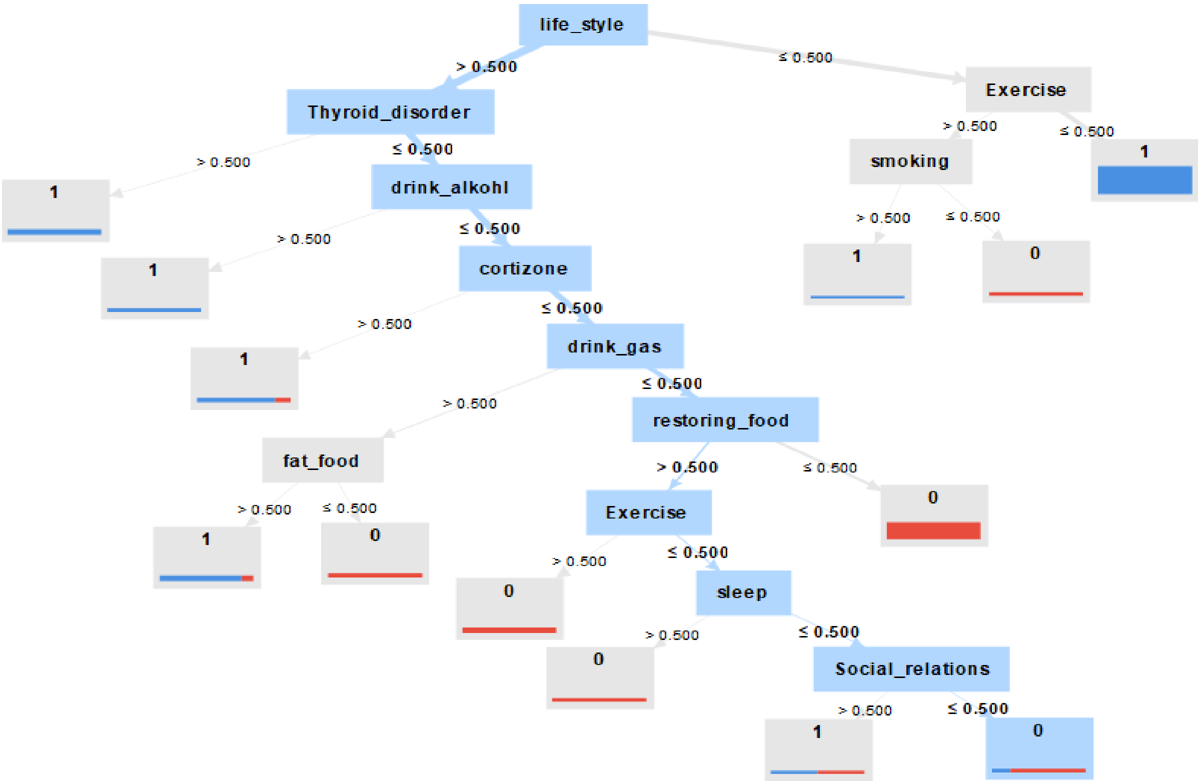


Figure 2. Draw of Decisions Tree Obesity Data Analysis for Student of Northern Technical University, Mosul, Iraq Data sample number 137

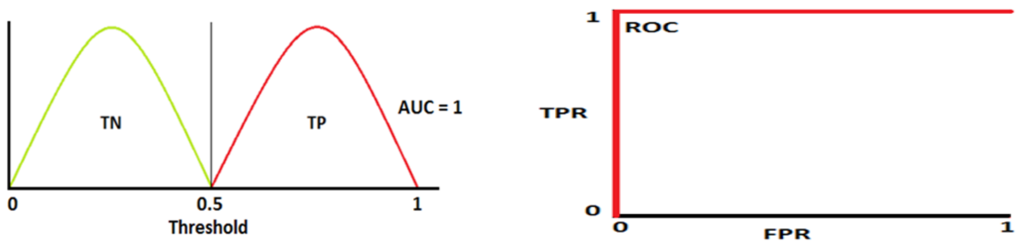


Figure 3. Distribution 1 Probability of Roc Curve

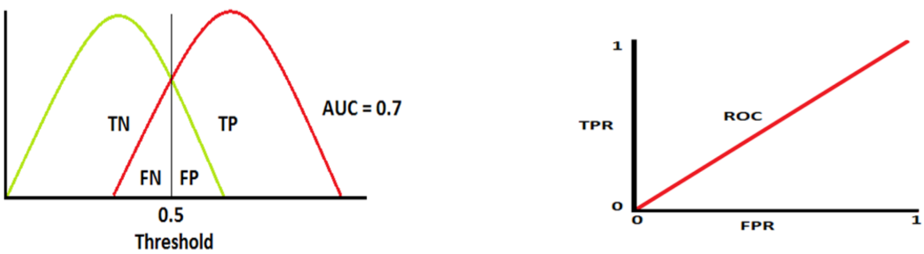


Figure 4. Distribution 2 Probability of Roc Curve

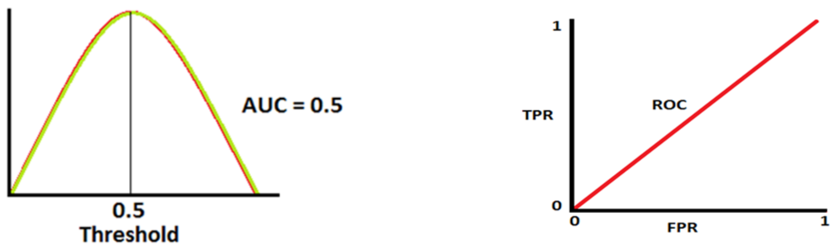


Figure 5. Distribution 3 Probability of Roc Curve

Questionnaire form for students of the Northern Technical University, medical departments, for factors related to obesity

Department: Specialization

Age: Gender:

Length: Weight: Waist measurement:

Dear student, please choose the appropriate answer by

checking the box with a sigh ✓

1- Deal with TV or Computer t	**: yes (1)**	**no (0)**
2- Do you exercise	**: yes (1)**	**no (0)**
3- Do you smoke	**: yes (1)**	**no (0)**
4- Is your sleep nature normal	**: yes (1)**	**no (0)**
5- Do you always eat junk food	**: yes (1)**	**no (0)**
6- Do you eat regularly?	**: yes (1)**	**no (0)**
7- Do you eat fast	**: yes (1)**	**no (0)**
8- Do you drink alcoholic beverages	**: yes (1)**	**no (0)**
9- Do you have soft drinks?	**: yes (1)**	**no (0)**
10- There is obesity in family history	**: yes (1)**	**no (0)**
11- Do you have a malfunction of the thyroid gland?	**: yes (1)**	**no (0)**
12-You always have the habit of resorting to food	**: yes (1)**	**no (0)**
13- Are you taking cortisone medications?	**: yes (1)**	**no (0)**
14- Are social relations good	**: yes (1)**	**no (0)**
15- Always eat fatty foods	**: yes (1)**	**no (0)**
16- Do you live an easy lifestyle?	**: yes (1)**	**no (0)**
17- Do you work abroad?	**: yes (1)**	**no (0)**
18- Obesity Index	**: Obese (1)**	**Normal (0)**
19- Gender	**: male (1)**	**female (0)**

We wish you all the best of luck and success

Investigator