IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Classification Algorithms for Prediction of Obesity Levels based on Food Consumption and Physical Conditions using WEKA

[1]A. Jabarali, [2]S. Baby Sandhya and [3]P. A. Vanithapriya

[1]Assistant Professor, [2,3]Post-graduate Student

[1]Department of Statistics

[1]Madras Christian College, East Tambaram, Chennai 600 059, Chengalpattu, Tamil Nadu, India.

Abstract: Obesity has been steadily growing from three decades and strongly associated with several disease. There are several factors has been used for diagnose the obesity and over-weights. In this study, to select, explore and model the data-set of obesity level with food habits and physical activities using classification algorithms. The high accuracy and low error measures of classification algorithm is identified to predict the obesity level based on performance and error measures using WEKA.

Index Terms **- Obesity, Prediction, Descriptive Measures, Chi-Square Test, Classification Algorithms, WEKA.**

I. INTRODUCTION

According to the World Health Organization (WHO) overweight and obesity are considered as the excessive accumulation of fat in the body which has been measured by using Body Mass Index (BMI). BMI is a ratio of weight of the person in kilograms and square of height in meters. A high BMI can be an indicator (between 25 to 30) as overweight and if it exceeds more than 30 it is known as the condition in which a person is obese which is a major risk for health of the individual. The risk of obesity has drone worldwide with almost Four million peoples dying since 2017 due to overweight and obesity according to the global burden of disease. Rates of overweight and obesity continue to grow in adults and children. From 1975 to 2016, the prevalence of overweight or obese children and adolescents aged 05-19 years increased more than four-fold from 4% to 18% globally. people may think obesity can be caused due to excessive consumption of food but is actually due to condition in which any individual is facing lack of Nutrition that is Malnutrition. This has been an issue only in the developed countries but now it is being prevalent more in the underdeveloped or developing countries where rate of increase has been 30% more than the developed countries. Obesity and overweight are also the major risk factors for many chronic disorders including cardiovascular disease such as heart disease or stroke, diabetes which is associated with conditions like blindness or limb amputation and many different types of cancers causing major deaths around the world. It may also lead to musculoskeletal disorders like osteoarthritis.

Shirin, et al., (2015) studied health impacts of obesity. They concluded that Overweight, obesity and their impacts in different dimensions of health must be considered as one of the most important public health priorities. Also, there is a need for comprehensive strategies for prevention and control of the obesity and overweight.

Kapil et al., (2018) studied in the obesity prediction using Ensemble Machine Learning Approaches. The machine learning prediction approach of leverages generalized linear model, random forest, and partial least squares were used. Also, the health parameters have been predicted and recommend corrective measures based on obesity values.

Correa et al., (2019) used the SEMMA data mining methodology to select explore and model the data set using Decision Tree, J48, Bayesian Networks, Naïve Bayes and Logistic Regression. The data was related to the young undergraduate students aged between 18 and 25 from Columbia Mexico and Peru with 712 records which include 324 and 388 women with various physical measures and food habits. They concluded that the best results were given by the decision tree J48 algorithm using the performance measures, precision (97.4%), recall (97.8%), true positive rate (97.8%) and false positive rate (0.2 %), improving the results obtained in the previous studies with similar background.

Ayan et al., (2020) studied on identification of risk factors associated with obesity and overweight with Machine Learning Techniques. They studied the performance of different machine learning algorithms of Artificial neural networks, random forests, Support vector machines, logistic regression. They identified model based on performance measures with risk factors.

Balbir (2020) studied on Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People. To classify adolescents of age 14 at a risk of becoming overweight or obese using BMIs measured at ages 3, 5, 7 and 11 years. The data is used from UK's Millennium Cohort Study (MCS), deals with the data pre-processing, application of ML algorithms for the classification of imbalanced data, techniques used for the treatment of data imbalance and the evaluation of suitable algorithms for improved prediction accuracy. The dataset is composed of 11110 instances and has three classes based on the obesity label. There are 8160 normal cases, 2126 overweight and 824 obese. Majority of the instances falling under the normal category, 19% under overweight and only just over 7% belonging to the obese category, makes the dataset highly imbalanced. Many machine learning algorithms has been used in order to predict the health conditions using a number of characteristic features. Issues relating to low prediction accuracy because of data imbalance have been considered and dealt with using Synthetic Minority Oversampling Technique (SMOTE). It is also proposed to use predict obesity state at the age of 14 using the obesity flags from earlier ages since the BMI is age and gender dependent for ages from 2 to 20.

Rodolfo and Ubaldo (2020) analyzed the obesity levels in 178 students which included 81 males and 97 females aged between 18 and 25 years studying in institutions in the countries of Colombia Mexico and Peru. comparative analysis on the classification algorithms which include decision tree support vector machine and simple k-means was performed using the evaluation mattresses of precision recall true positive and false positive rate and ROC area in WEKA after a preparation and transformation of the data to identify missing a typical data and correlation analysis. The result obtained by decision trees and simple k-means and precision (98.5%), recall (98.5%), true positive rate (98.5%), false positive rate (0.2%) and ROC area (99.5%) surface the results obtained in the previous studies, that had the precision level of 75% and 85%. The results obtained in this study can be used to analyses the patterns of various pathological diseases, detect them to the earliest as possible and minimize the impact of the disease based on computational intelligence.

Joao and Arise (2020) studied the association between obesity and higher levels of hospitalization 4 outcomes and mortality due to COVID-19 which included studies published between December 2019 to May 2020. The research contains inclusive criteria targeting studies of human's adults infected by SARS-COV-2, with or without comorbidities. 20 articles were included with the population that is estimated between 1 to 1761 auto switch at least four cases are found to have association between the obesity and mortality rate and 85.3% where hospitalized and 14 of them was found to have more complications in covid-19 due to obesity. The vault was concluded that persons with obesity have vagrant condition of COVID-19 and increase in prevalence of hospitalization when occurring with other chronic condition in the individual and elderly individuals as well.

The objective of the paper is to study the different classification algorithms, the association between the obesity level with food habits and physical activities, compare the different classification algorithms based on accuracy measures and predict the obesity levels in individuals based on the Physical activities and Food habits.

This paper is organized with four sections and bibliography. Section 1 includes the introduction about the obesity, recent related works on obesity and the objectives of the study. Section 2 describes the methodology for which the data to be analysed in Section 3. The results of the statistical data analysis are presented in Section 3. Section 4 contains the conclusion of the study which are made with respect to objectives. Finally, the bibliography of the to carry out this work is attached.

II. METHODOLOGY

This section discusses the descriptive measures, chi-square and various classification algorithms. The performance measures and error measures of classification algorithm are presented.

2.1. Descriptive Measures

A descriptive measure is a summary statistic that quantitatively describes or summarizes the collection of information, while descriptive statistics is the process of using and analysing those statistics. Some measures that are commonly used to describe a data set are measures of central tendency, measures of variability or dispersion and kurtosis and skewness. Measures of central tendency include the mean, median, mode and quantiles; measures of variability include the standard deviation (or variance), the minimum and maximum values of the variables.

2.2. Chi-Square Test for Association of Attributes

The association of attribute is measured by the degree of relationship of two phenomena, whose sizes are not directly measurable but studied by the presence of a particular attribute. The Chi-square test for association of attributes is used to determine if there is any association between the two attributes. The hypothesis and test statistic are as follows;

H_0: There is no significant difference between two attributes ie., The two variables are not associated (Independent)

H_1: There is no significant difference between two attributes ie., The two variables are associated (Related or dependent)

and $\chi^2 = \sum \frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$

2.3 Classification Algorithms

Classification is a supervised machine learning technique which helps in classifying a data into relevant previously learnt categories. The two steps involved in classification are the categorized and labelled training data is fed into the system to develop understanding of the different categories based on the understanding developed from the training data the algorithm will classify similar data for the unlabeled data detecting fraud regarding email spam is the common application of the classification technique. It is to be noted that the classification can be performed for two or more categories in a training data which is classified into categories is fit into the system for the understanding of the categories and then testing data which is unlabeled is fit into the system which it classifies itself.

The performance measures are Accuracy, Precision, Recall or Sensitivity, F-Measure, Kappa Statistic. For classifier algorithms these measures are the percentage of records in the test dataset that are correctly predicted by the classifier. To compute all the above-mentioned measures, a useful tool is used to analyses how well a classifier predicts different classes is Confusion Matrix. A confusion matrix is a table that helps to study about the performance of the classifier and tells the number of predictions(cases) used to identify on the test dataset for which true values are known. In the confusion matrix, the rows represent the predicted class, and the columns represents the actual class as shown in Table 1.

Table 1. Confusion Matrix

		Actual Class	
		Genuine (P)	Counterfeit (N)
Predicted Class	Genuine (P)	TP	FP
	Counterfeit (N)	FN	TN

where, True Positive (TP) is the number of predictions where the classifier predicts the positive class as positive correctly; True Negative (TN) is the number of predictions where the classifier predicts the negative class as negative correctly; False Positive (FP) is the number of predictions where the classifier predicts the negative class as positive incorrectly; False Negative (FN) is the number of predictions where the classifier predicts the positive class as negative incorrectly.

Accuracy is the number of correctly classified instances to the total number of predicted instances. Accuracy for the classifier is given as,

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP}$$

Precision is the proportion of the true positive (TP) against all the positive results including false positives. Precision for the classifier is gives as,

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall or sensitivity is a true positive (TP) rate. This is the proportion of positive results that are correctly identified. TP is the true positive rate, TN is true negative rate, FP is false positive rate and FN is false negative rate. Recall for the classifier is given as,

$$\text{Recall} = \frac{TP}{TP+FN}$$

F- Measure is a measure that combines precision and recall into a single measure.

$$\text{F-Score} = \frac{2TP}{2TP+FP+FN} = \frac{2*(Precision*Recall)}{Precision+Recall}$$

Kappa Statistic is used to measure the agreement between predicted and observed categorizations of a dataset. It is a chance corrected measure that ranges from -1 to 1. If kappa statistic is **"1",** then it indicates perfect agreement between predicted and observed category, and **"-1"** indicates perfect disagreement between predicted and observed category. If Kappa statistic is **"0",** it indicates there is no agreement, beyond which can be expected from chance. Kappa statistic is calculated by

$$\text{Kappa} = \frac{P(A)-P(E)}{1-P(E)}$$

Here,

P(A) = The Percentage agreement between classifier and the accuracy.

P(E) = The Chance agreement is given by,

$$P(E) = \frac{(TN+FP)(TN+FN)+(FP+TP)(FN+TP)}{(TN+FP+FN+TP)^2}$$

The error (or disturbance) is the amount by which an observation differs from its expected value, the latter being based on the whole population from which the statistical unit was chosen randomly. In statistics and optimization, errors and residuals are two closely related and easily confused measures of the deviation of an observed value of an element of a statistical sample from its "theoretical value". The error of an observed value is the deviation of the observed value from the (unobservable) true value of a quantity of interest (for example, a population mean), and the residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, a sample mean). Some commonly used measures of errors are given below:

Mean Absolute Error is the mean absolute difference between the estimated values of the variable and the actual values as observed during the study, ie. mean of the absolute deviations of variables from the forecasted values and is given by,

$$MAE = \sum_t \frac{|Y_t - \hat{Y}_t|}{N}$$

The root mean square error is the standard deviation of the residuals, ie., the standard deviation of the difference between the estimated and the observed value of the variable and is given by,

$$RMSE = \sqrt{\frac{\sum_t (Y_t - \hat{Y}_t)^2}{N}}$$

The measure of accuracy as a percentage can be found using the mean absolute percentage error. This is a more accurate measure as unlike other measures it expresses the error as a percentage and is given by

$$MAPE = \frac{\sum_t \left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|}{N} \times 100$$

Relative Mean Squared Error is the relative to what it would have been if a simple predictor has been used. It is the total squared error and normalizes it by dividing by the total squared error. By taking the root of the relative mean squared error reduces the error to the same dimensions as the quantity being predicted. It is the measure of accuracy as a percentage. It can be expressed as,

$$E_i = MSPE = \sqrt{\frac{\sum_j (P_{ij} - T_j)^2}{\sum_j (T_j - \bar{T})^2}}$$

III. Analysis of Obesity Levels Based on Food Consumption and Physical Conditions

This section presents the data description about the obesity level. The descriptive measures of the attributes in the data were discussed. The appropriate classification algorithm has identified for predict the obesity level through various accuracy measures like F-measure, Precision, Accuracy, Kappa statistic, along with the Mean Absolute Error (MAE), Relative Mean Absolute Error (RMAE), Root Mean Square Error (RMSE) and Root Relative Squared Error (RRSE) in comparison with various classification algorithms.

3.1. Data Description

The secondary data is obtained from UCI repository for machine learning which contains Four Hundred and Ninety-Eight (498, in number) observations with Sixteen (16, in number) attributes which includes Three (3, in number) scale variables and Thirteen (13, in number) categorical variables (Table 2). The categorical and scale variables of obesity levels along with possible values has been given in Table 2. The output categorical variable of obesity levels has classified into Seven (07, in number) classes namely Insufficient Weight, Normal Weight, Overweight Level-I, Overweight Level-II, Obesity Level-I, Obesity Level-II and Obesity Level-III.

3.2. Descriptive Measures of Obesity Level

The frequency measures for categorical and scale variables are presented in APPENDIX. The association between the obesity levels with all the other qualitative attributes are tested using Chi-square test statistic.

The various descriptive measures of the quantitative variables, age, height and weight, of the data which includes measures of central tendency like mean, median, mode and quartiles and measures of dispersion like range, standard deviation, variance, skewness, and kurtosis are computed and tabulated in the below Table 3. Based on mean, median and mode, the scale variables of obesity followed positively skewed distribution since the median and mode would be to the left of the mean. ie., means that the mean is greater than the median and the median is greater than the mode (Mean > Median > Mode).

3.3. Chi-Square Test for Association Between Attributes

The Association between the Obesity levels with all the other qualitative attributes of the data are tested and presented in Table 4. It is found that, the attributes; Family history with overweight, Smoking, Number of main meals and the usage of technological devices are found to be significant to the Obesity levels of the individuals and the rest of the seven qualitative attributes are found to be non-significant to Obesity levels.

3.4. Classification Algorithms for Obesity Level

The dataset was spited into two different sets including training and test sets. There are Four Hundred and Ninety-Eight (498, in number) instances which spilt into training set and test set without overlapping that means the set will not contain duplicate items. In training set, it contains 75% of data (i.e., 374) instances and the rest of the 25% of data (i.e., 124) are considered as cross validation and test data. After training the data, test the data to the training set. At the end, the result file shows the predicted results with the help of WEKA software. The WEKA output of confusion matrix of different classifiers is given Table 5.

Table 2. Description of the Dataset on Obesity Levels

Attributes	Name	Possible Values	Types
Gender	Gender	Male, Female	Nominal
Age	Age	Age in Years	Scale
Height	Height	Height in Met	Scale
Weight	Weight	Weight in Kilograms	Scale
Family History with Overweight	Family History with Overweight	Yes, No	Nominal
FAVC	Frequent Consumption of High Caloric Food	Yes, No	Nominal
FCVC	Frequent Consumption of Vegetables	Never, Sometimes, Always	Nominal
NCP	Number of Main Meals	Between 1 and 2, Three, More than three	Nominal
CAEC	Consumption of Food between Meals	Always, Frequently, Sometimes, Never	Nominal
SMOKE	Smoke	Yes, No	Nominal
CH2O	Consumption of Water Daily	Less than a liter, Between 1 and 2 liters, more than 2 liters.	Nominal
SCC	Calories Consumption Monitoring	Yes, No	Nominal
FAF	Physical Activity	I don't have, 1 or 2 days, 2 or 4 days, 4 or 5 days	Nominal
TUE	Time using Technology Devices	0-2 hours, 3-5 hours, More than 5 hours	Nominal
CALC	Consumption of Alcohol	Always, Frequently, Sometimes, Never	Nominal
MTRANS	Transport Used	Automobile, Bike, Motorbike, Public Transport, Walking	Nominal

From Table 6 it is observed that, the J48 classifier is the best classifier for predicting the obesity level based on food consumption and physical conditions since it has minimum error and maximum precision in compared with other classifiers. The tree diagram of J48 classifier has presented in Figure 1. By using this condition of nodes and leaves of tree, the future individuals have classified.

Table 3. Descriptive Measures of Heights, Weights and Ages of Individuals

		Ages	Heights	Weights
Mean		23.147	1.686	69.57
Median		21	1.680	67
Mode		18	1.7	60
Std. Deviation		6.722	0.098	17.013
Variance		45.179	0.009	289.449
Skewness		2.202	0.201	1.186
Kurtosis		6.070	-0.519	2.938
Range		47	0.53	134
Minimum		14	1.450	39
Maximum		61	1.980	173
Percentiles	25	19	1.613	58
	50	21	1.680	67
	75	24	1.750	80

Table 4. Association between Obesity Levels - wise - Physical and Food Habits

Attributes	Test Statistic Value	Asymp. Sig. (Two-Sided)	Result
Obesity levels vs. Family History with Overweight	28.200	0.0001*	Significant
Obesity levels vs. Smoke	17.724	0.007*	Significant
Obesity levels vs. Frequent Consumption of Vegetables	6.1050	0.4115	Non-Significant
Obesity levels vs. Frequent Consumption of High Caloric Food	14.451	0.2729	Non-Significant
Obesity levels vs. Number of Main meals	24.323	0.0184*	Significant
Obesity levels vs. Consumption of food between meals	25.466	0.1126	Non-Significant
Obesity levels vs. Calories Consumption monitoring	10.739	0.0968	Non-Significant
Obesity levels vs. Physical Activity	23.879	0.1591	Non-Significant
Obesity levels vs. Usage of Technology devices	21.931	0.0383*	Significant
Obesity levels vs. Consumption of Alcohol	16.896	0.5302	Non-Significant
Obesity levels vs. Mode of Transport	36.298	0.0513	Non-Significant

Table 5. 2×2 Confusion Matrix for Different Classifiers

S. No.	Classifiers	Confusion Matrix
1.	Naïve Byes	$\begin{bmatrix} 78 & 46 \\ 46 & 0 \end{bmatrix}$
2.	Random Forest	$\begin{bmatrix} 85 & 39 \\ 39 & 0 \end{bmatrix}$
3.	J48	$\begin{bmatrix} 103 & 21 \\ 21 & 0 \end{bmatrix}$
4.	Logistic Regression	$\begin{bmatrix} 100 & 24 \\ 24 & 0 \end{bmatrix}$
5.	Neural Networks	$\begin{bmatrix} 77 & 47 \\ 47 & 0 \end{bmatrix}$
6.	K-Means Nearest Neighbourhood	$\begin{bmatrix} 53 & 71 \\ 71 & 0 \end{bmatrix}$

Table 6. Classifier's Performance on the Basis of Classified Instances and Error Measures

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Precision	Recall	F - Measure	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Naïve Bayes	78 (62.90%)	46 (37.09%)	0.4588	0.6290	0.6290	0.6290	0.3469	0.1134	0.2839	62.52%	94.07%
Random Forest	85 (68.55%)	39 (31.45%)	0.5215	0.6855	0.6855	0.6855	0.4472	0.1204	0.2367	66.38%	78.41%
J48	**103(83.06%)**	**21 (16.94%)**	**0.7103**	**0.8306**	**0.8306**	**0.8306**	**0.7281**	**0.0501**	**0.1994**	**27.64%**	**66.09%**
Logistic Regression	100(80.64%)	24 (19.35%)	0.6757	0.8065	0.8065	0.8065	0.6901	0.0560	0.2320	30.86%	76.88%
Neural Networks	77 (62.09%)	47 (37.90%)	0.4503	0.6210	0.6210	0.6210	0.3581	0.1251	0.2727	68.96%	90.35%
K Nearest Neighbour	53 (42.74%)	71 (57.26%)	0.2718	0.4274	0.4274	0.4274	0.0983	0.1651	0.4008	91.02%	132.79%

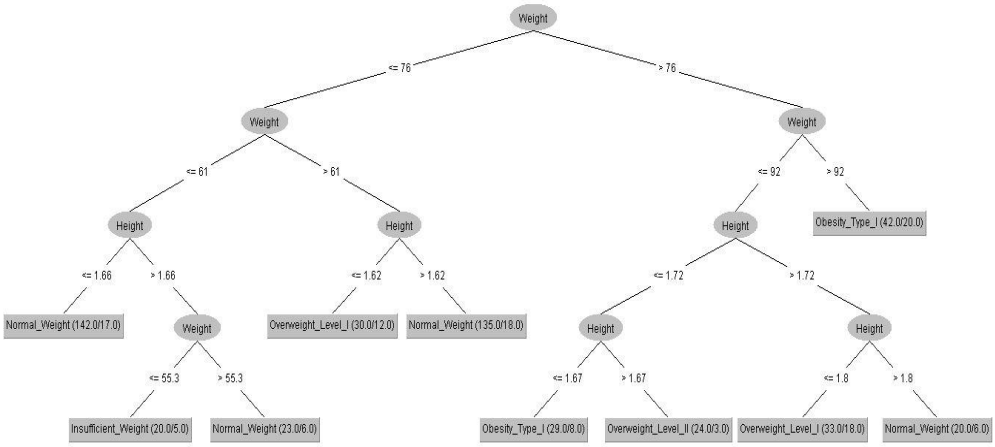

Figure 1: Tree Diagram for J48 Classifier

IV. SUMMARY AND CONCLUSION

In data mining, classification algorithms are used frequently for efficient decision making. It is one of the data mining tasks that classify an instance into the target attributes or class which are mutually exclusive and exhaustive and also help in predicting the target class of the future dataset. This study identifies the best classifier amongst the six classifiers for the estimation of obesity levels based on physical activities and food consumption. Before, identification of the model, the frequency analysis for categorical variables and descriptive measures for scale variables has done. From descriptive measures of scale variables, age, height, weight is positively skewed distributed. Meanwhile, the association between the obesity levels with all the other categorical variables are tested and it is found that the attributes, Family History with Overweight, Smoking, Number of Main Meals and the Usage of Technological Devices are found to be significant to the Obesity levels of the individuals and the rest of the seven categorical variables are found to be non-significant to obesity levels.

Based on the model procedure proposed, the obesity level data set was applied to be six supervised machine learning algorithms in WEKA and the confusion matrix along with the accuracy measures are obtained for each classifier. Here, the J48 decision tree was identified to be the best classifier amongst the other proposed classifiers in terms of accuracy and precision along with minimum error. Although, other classifiers have reasonable accuracy and precision, from the study the J48 decision tree is suggested to be the best classifier for the estimation of obesity level best on physical activity and food consumption data set to give higher accuracy and precision. Hence, the best prediction could be obtained from J48 classifier for the diagnosis of obesity level of future individuals based on the Sixteen (16, in number) attributes which includes Three (3, in number) scale variables and Thirteen (13, in number) categorical variables.

REFERENCES

[1] Abdullah, F. S., Manan, N.S.A., Ahmad, A., Wafa, S.W., Shahril, M.R., Zulaily, N. and Ahmed, A. (2016). Data Mining Techniques for Classification of Childhood Obesity among six-year school children", International Conference on Soft Computing and Data Mining, Springer, 549, 415-474.

[2] Adnan, M.H.M. and Hussain. W. (2011). A Framework for Childhood Obesity Classifications and Predictions Using NB Tree, Information Technology in Asia (CITA 11) 2011, 7th International Conference, IEEE, 1-6.

[3] Adnan, M.H.M. and Hussain. W. (2011). A Hybrid approach using Naïve Bayes and Genetic Algorithm for Childhood Obesity Prediction, International Conference Computer and Information Science (ICCIS), IEEE,1, 281-285.

[4] Ayan, C., Martin, W. G., and Martinez, S. G. (2020). Identification of Risk Factors Associated with Obesity and Overweight - A Machine Learning Overview, Multidisciplinary Digital Publishing Institute, 20(9), 2734.

[5] Balbir, S. H. T. (2020). Early Prediction of the Risk of Overweight and Obesity in Young People, Computational Science-ICCS 2020, Nature Public Health Emergency Collection, 12140, 523-535.

[6] Correa, E. D., Fabio, E. M., Alexis, D., Roberto, C. M. and Sánchez, H. B. A. (2019). Obesity Level Estimation Software based on Decision Trees, Journal of Computer Science, 15(1), 67-77.

[7] Dugan, T.M., Mukhopadhyay, S., Carroll, A. and Downs, S. (2015). Machine Learning Techniques for Predictions of Early Childhood Obesity, Applied Clinical Informatics, 6(3), 506-520.

[8] Han, J. and Kamber, M. (2006). Data Mining Concepts and Techniques. San Francisco, CA: Morgan Kaufmann.

[9] Joao, V. V. and Arise, G. G. (2020). Impact of Obesity on Hospitalizations and Mortality due to Covid-19: A systematic review, Obesity Research & Clinical Practice, 14(5), 398-403.

[10] Kapil. J., Niyati, B. and Prashant, S. R. (2018). Obesity Prediction using Ensemble Machine LearningApproaches, Proceedings of the 5th ICACNI 2017, Recent Findings in Intelligent Computing Techniques, 2, 355-362.
[11] Manna, S. and Jewkes, A.M. (2014). Understanding Early Childhood Obesity Risks: An Empirical Study using Fuzzy Signatures, In fuzzy Systems (FUZZ –TREE) 2014 IEEE, International Conference, 1333-1339.
[12] Michie, D., Spigelhater, D. J., Taylor, C. C. and Campbell, J. (1994). Machine Learning, Neural and Statistical Classification. Upper Saddle River Ellis Horwood.
[13] Moody, A. (2013). Adult Anthropometric Measurements Overweight and Obesity, Health and Social Care Information Center.
[14] Payan, C. D., DeGuzman, M., Johnson, K., Serban, N., and Swann, J. (2015). Estimating Prevalence of Overweight or Obese Children and Adolescents in small Geographic Areas Using publicly available data, Preventing Chronic Disease, Centres for Disease Control and Prevention, 12 (140229), E32.
[15] Rodolfo, C. C. and Ubaldo, M. (2020). Estimation of Obesity Levels Based on Computational Intelligence, Informatics in Medicine Unlocked, 21, 100472.
[16] Rohit. A., and Suman (2012). Comparative Analysis of Classification Algorithms on Different Datasets using WEKA, International Journal of Computer Applications, 54(13), 21-25.
[17] Rokholm, B., Baker, J.L. and Sorensen, T. (2010). The Levelling off of the Obesity Epidemic Since Year 1999- A Review of Evidence and Perspectives, International Association for the Study of Obesity, 11, 835-846.
[18] Shirin, D., Mostafa, Q., Niloofar, P. and Roya, K. (2015). Health impacts of obesity, Pakistan Journal of Medical Sciences, Professional Medical Publications, 31(1), 239-242.
[19] Suguna, M. (2016). Childhood Obesity Epidemic Analysis Using Classification Algorithms, International Journal of Modern Education and Computer Science, 4(1), 22-26.
[20] Zhang, M.L. and Zhore, Z.H. (2009). Multi-instance Clustering with applications to Multi-instance Predictions, Applied Intelligence, 31(1), 47-68.

Appendix -A: Frequency Measures of Obesity Attributes

ATTRIBUTES	POSSIBLE VALUES	INDIVIDUALS (%)	ATTRIBUTES	POSSIBLE VALUES	INDIVIDUALS (%)
Gender	Male	271 (54.42%)		More than three	46(9.23%)
	Female	227 (45.58%)		**Total**	**498 (100%)**
	Total	**498 (100%)**	**Consumption of Food between Meals**	Always	53(10.64%)
Age (in Years)	Child	0 (0%)		Frequently	136(27.31%)
	Adolescence	38 (7.63%)		Never	20(4.02%)
	Adult	459 (92.17%)		Sometimes	289(58.03%)
	Senior Adult	01 (0.20%)		**Total**	**498 (100%)**
	Total	**498 (100%)**	**Consumption of Water (in litres)**	Less than a litre	135 (27.11%)
Heights (in Metres)	1.40 – 1.50	02 (0.4%)		Between one and two litres	266(53.41%)
	1.50 – 1.60	79 (15.86 %)		More than two litres	97 (19.48%)
	1.60 – 1.70	183 (36.75%)		**Total**	**498 (100%)**
	1.70 – 1.80	151 (30.32%)	**Monitoring Calorie Consumption**	Yes	55(11.04%)
	1.80 – 1.90	75 (15.06%)		No	443(88.96%)
	1.90 – 2.00	08 (1.61%)		**Total**	**498(100%)**
	Total	**498 (100%)**	**Physical Activities**	No	162(32.53%)
Weights (in Kilograms)	30 – 40	1(%0.20)		one or two days	158(31.73%)
	40 – 50	33(6.62%)		two or four days	113(22.69%)
	50 – 60	114(22.89%)		four or five days	65(13.05%)
	60 – 70	127(25.50%)		**Total**	**498(100%)**
	70 – 80	93(18.67%)	**Time spent on Technology Devices (in Hour)**	Less than 2	243(48.79%)
	80 – 90	75(15.06%)		Three to Five	181(36.35%)
	90 – 100	27(5.42%)		More than five	74(14.86%)
	100 – 110	13(2.61%)		**Total**	**498(100%)**
	110 – 120	8(1.61%)	**Frequency of Consumption of Alcohol**	Always	01(0.20%)
	120 – 130	5(1.00%)		Frequently	45(9.04%)
	130 – 140	1(0.20%)		Never	179(35.94%)
	>140	1(0.20%)		Sometimes	273(54.82%)
	Total	**498 (100%)**		**Total**	**498(100%)**
Family History with Overweight	Yes	300(60.24%)	**Mode of Transportation**	Automobile	99(19.88%)
	No	198(39.76%)		Bike	07(1.41%)
	Total	**498 (100%)**		Motor Bike	11(2.21%)
Smoking Habits	Yes	32(6.42%)		Public Transportation	326(65.46%)
	No	466(93.58)		Walking	55(11.04%)
	Total	**498 (100%)**		**Total**	**498(100%)**
Consumption of High Caloric Food	Yes	348(69.88%)	**BMI**	Insufficient Weight	34(6.83%)
	No	150(30.12%)		Normal Weight	287(57.63%)
	Total	**498 (100%)**		Overweight Level I	58(11.65%)
Frequent Consumption of Vegetables	Never	32(6.43%)		Overweight Level II	58(11.65%)
	Sometimes	272(54.62%)		Obesity Type I	47(9.44%)
	Always	194(38.95%)		Obesity Type II	11(2.21%)
	Total	**498 (100%)**		Obesity Type III	03(0.60%)
Number of Main Meals	Between one and two	108(21.69%)		**Total**	**498(100%)**
	Three	344(69.08%)			