ISSN No:-2456-2165

Five Machine Learning Supervised Algorithms for The Analysis and the Prediction of Obesity

Jakin KABONGO
Senior Lecturer at the University of Kinshasa/Department of Computer Science and Business English. MS in ICT Convergence at Handong Global University, Pohang/Republic of Korea

MPOVO LUZOLO CLEM'S
Master of Agriculture in Food Security and Agriculture Development, Korea National University-KNU, School of Applied Biosciences/Crop Physiology; Senior Lecturer at ISP POPOKABAKA/D.R Congo

PAVODI MANIAMFU
PhD Student, University of Tsukuba/JAPAN in Deep Learning and Epidemiological Models. Junior Lecturer at The University of Kinshasa/DR Congo.

DUMBI KAMANDA LOUISON
MA in Environmental Communication; University of Kinshasa/D.R. Congo; Junior Lecturer at The University of Kinshasa

Abstract:- Obesity and overweight are major risk factors for a variety of chronic diseases, including cardiovascular diseases like heart disease and stroke, which are the main leading causes of most deaths worldwide. Obesity can also lead to diabetes and its complications, such as blindness, limb amputations, and the need for dialysis. Diabetes prevalence has quadrupled worldwide since 1980. Excess weight can also cause musculoskeletal disorders such as osteoarthritis. The objective of this research is to analyze and predict obesity using machine learning algorithms to assist clinicians and public health agents to make an optical decision related to the prevention, the detection, and the treatment of obesity. Five machine learning classification models including Random Forest, Support Vector Machine, Logistic regression, K-nearest Neighbor, and Ridge Classifier were used for the purpose. These five models were trained after the Exploratory Data Analysis and the Data Preprocessing with k-ford cross-validation, classification report, the confusion matrix, and the learning curve as metrics. After the training according to the accuracy performance given by each model and the learning curve, the Support Vector Machine was selected and optimized as the final model with 97% of accuracy.

Keywords:- *Obesity, Machine-Learning Algorithms, Premature Deaths.*

I. INTRODUCTION

Obesity is among the leading causes of premature deaths in the world [1] [2]. According to research carried out by the Global Burden of Disease, obesity was responsible for about 4.7 million deaths globally in 2017[1]. The World Health Organization evaluated this number as about four times of individuals who died in car accidents, and nearly five times those who died from HIV/AIDS in 2017[1]. Obesity is nowadays regarded as one of the world's most serious public health issues. It is generally used as a metric to detect changes in a population's health and nutrition [1][3][4]. This is crucial in public health because there is solid evidence that obesity increases the risk of a variety of unfavorable consequences, including cardiovascular diseases with 17.9 million annual deaths, diabetes with 1.6 million annual deaths, cancers with 9.0 million annual deaths, and respiratory diseases with 3.9 million of annual deaths, to name but a few [1]. These four diseases are the main components of non-communicable diseases, which is responsible for 41 million deaths each year accounting for 71% of all the deaths worldwide and 15 million of these deaths occur prematurely (between the age of 30 to 69 years old)[5][6]. 85% of these premature deaths occur in low-and-middle-income countries where Vulnerable as well as socially disadvantaged people get sicker and die younger than those in higher social positions, owing to increased exposure to harmful products such as tobacco or unhealthy dietary habits, and also limited health service access [5].

Obesity and overweight are caused by an energy imbalance between total calories consumed and calories expended. According to WHO there has been worldwide an increase in the consumption of energy-dense foods high in fat and sugar content; and an increase in physical inactivity as a result of social and environmental changes related to development, as well as a lack of supportive policy initiatives in sectors such as health, agriculture, transportation, urban planning, environment, food processing, distribution, marketing, and education that frequently cause changes in dietary and physical activity patterns[5][2].

Once thought to be a problem only in high-income countries, Overweight and Obesity are already on the rise in low- and middle-income nations due to the consumption of processed food (high-salt, high-fat, high-sugar, energy-dense, and micronutrient-deficient), most of them from western countries combined with a lower level of physical activity, sedentary lifestyle, and the excessive use of technological appliances[5]. Low-and-middle-income countries are nowadays undergoing a double burden with the coexistence of undernutrition and overnutrition in the same population while these countries continue to struggle with infectious diseases (communicable diseases) they are also undergoing a sensible increase in non-communicable diseases(NCDs), particularly in urban areas[7][5][8]. Since

ISSN No:-2456-2165

2000, the number of overweight children under the age of five has increased by nearly 24% in Africa [7][8]. In 2019, Asia was home to nearly half of all children under the age of five who were overweight or obese [2].

The Healthcare industry has long been one of the significant beneficiaries and adopters of Information and Communication Technological (ICT) advancements [9][10]. Machine learning as a subset of artificial intelligence is now used in a variety of health-related fields, including the development of new medical procedures, the management of patient data and records, and the diagnosis and treatment of diseases [9]. Therefore, this research employs machine learning techniques to analyze and predict obesity based on nutritional habits, physical activity, and other characteristics to assist clinicians and public health agents particularly those in low-and-middle-income countries in intervening sooner and quicker as said Sebastian Thrum to New Yorkers "just as machines have made human muscles thousand times more powerful, machines will make the human brain a thousand times more powerful" (wikipedia.org/wiki/Sebastian_Thrun).

II. PREVIOUS STUDIES

Since the rise of non-communicable diseases that coincided with the breakthroughs in sciences, especially in the field of Information and Communication Technologies (ICT), several pieces of research have been carried out to propose ICT-based solutions to assist clinicians and public health agents to make optimal decisions.

Concerning obesity, several solutions have been already studied as revealed in this systematic literature review [11] carried out to explore obesity research and machine learning strategies for treatment and prevention of obesity from 2010 to 2020 to identify machine learning algorithms that can be utilized to better predict obesity. As a result, from an initial pool of over 700 papers on obesity, 93 papers were recognized as primary research from the review articles. This literature was performed to assist decision-makers in better understanding the impact of obesity on public health and identify outcomes that could be used to assist health authorities as well as public health officials in further mitigating threats and effectively guiding obese people around the world [11].

Several papers have used a multi-algorithm approach to predict obesity to select the one with the highest performance as their final model. For example, [12] Used nine machine learning algorithms to predict obesity (random forest, multilayer perceptron, k-nearest neighbor, support vector machine, adaptive boosting, logistic regression, nave Bayes, gradient boosting classifiers, and decision trees). The Logistic Regression was the model with the best performance accuracy of 97.09 percent. [13] Used five data mining supervised and unsupervised techniques to create a computational intelligence to identify obesity levels based on lifestyle including, the Light Gradient Boosting Machine (Light GBM) classifier, random forest (RF), decision tree (DT), Extremely Randomized Trees (ET), and logistic regression (LR). The best performance was given to the LightGBM classification model with a performance accuracy of 0.9990. [14] Used six machine learning algorithms: Decision Tree, Random Forest, ID3, Nave Bayes, Bayes, and J48 to predict obesity in a child after having reached two years old. Finally, ID3 got the best score with 85% of accuracy and 89% of sensitivity. [15] Employed four data mining classification techniques to predict obesity (Logistic Model Tree-LMT, Random Forest-RT, Multi-Layer Perceptron-MLP, and Support Vector Machines-SVM). The Logistic Model Tree performed the best in terms of precision, with a score of 96.65%.

A. Data Collection

Data from the UCI (University California Irvine) Machine Learning Repository, which is a Dataset for evaluating obesity levels in patients based on their dietary habits and physical activity [17]. This dataset contains 17 variables with 2111 records labeled with the class variable “NObeyesdad” (Obesity Level) [17]. The Exploratory Data Analysis in this study is carried out according to data delivered by the Global Burden of Disease (GBDx), the World Health Organization reports (WHO), and the Institute of Metrics and Evaluation (IHME) in what concerning obesity and its collateral effects.

B. Exploratory Data Analysis

At this stage, we critically investigated the dataset to gain an exhaustive understanding of variables, patterns, data anomalies, and correlations between variables to design a policy that leads to a reliable prediction of obesity.

➢ *Visualization of the Dataset*

The dataset contains seventeen variables represented in columns and 2,111 records represented in rows of which 52.9% are qualitative (object) and 47.1% of variables are numerical (float64). The Target variable is NObeyesdad, which is a multiclass variable indicating whether a person(patient) is normal weight, insufficient weight, obesity type I, obesity type II, obesity type III, overweight level I, and overweight level II.

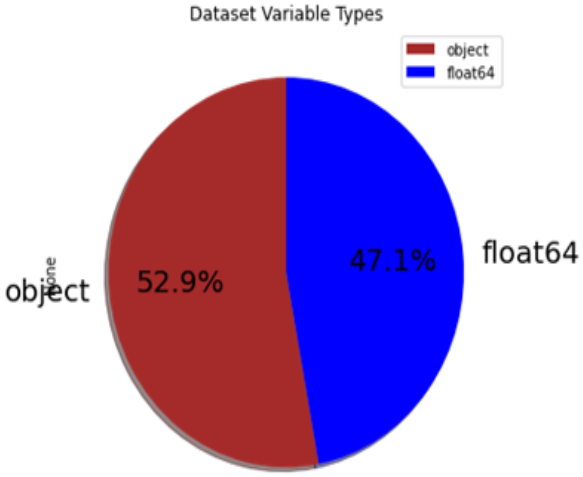

Fig 1 Dataset Variable Types

ISSN No:-2456-2165

C. Target Variable (Level Of Obesity)

The analysis of the target variable shows that 551 cases or 16.6% belong to the class category of ‘’Obesity Type I’’, 324 cases or 15.3% belong to the class category of “Obesity Type III”, 297 cases or 14% belong to the class category of “Obesity Type II”, 290 cases or 13.7% belong to the class category of ‘’Overweight Level II, 290 cases or 13.7% belong to the class category of "Overweight Level I", 287 cases or 13.5% belong to the class category of "Normal Weight, and finally, 272 cases or 12.8% belong to the class of "Insufficient Weight". This result shows that there is quite an unbalance in the repartition of class categories as presented in the figure below.

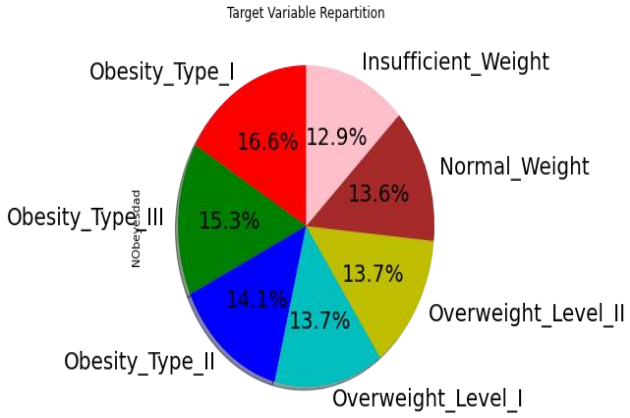

Fig 2 Target Variable Class Categories

D. Independent Variables Analysis (Inputs)

This dataset contains sixteen feature variables (X) grouped in three subsets. The first subset contains features related to ‘eating habits in which there are the following variables: ‘Frequency of consumption of vegetables’ (FCVC), ‘Frequent consumption of high caloric food’ (FAVC), ‘Number of main meals’ (NCP), ‘Consumption of food between meals’ (CAEC), ‘Consumption of alcohol’ (CALC), and ‘Consumption of water daily’ (CH20). The second subset contains feature variables related to the physical condition with variables such as: ‘Physical activity frequency’ (FAF), ‘Calories consumption monitoring’ (SCC), ‘Transportation used’ (MTRANS), and ‘Time using technology devices’ (TUE). The third subset contains additional attributes characterizing individuals according to their ‘Age’, ‘Gender’, ‘Weight’, and ‘Height’. The table below illustrates the summary of each feature variable within the dataset.

Table 1 Summary of the Dataset Independent Variables

S No.	VARIABLES	DESCRIPTION	VARIABLE SUBSET	TYPE
1	**Gender**	Male: 1068 (50.59%) Female: 1043 (49.41%)	Characteristics	Object
2	**Age**	Max:61 Min:14 Mean: 24.3 Std: 6.3	Characteristics	Float
3	**Height (**Height of a person in meters)	Max: 1.98 m Min: 1.45 m Mean: 1.7 m Std: 0.093 m	Characteristics	Float
4	**Weight (**Weight of a person in kilograms)	**Max**: 173 kg **Min**: 39 kg **Mean**: 85.5 kg **Std**: 26.1 kg	Characteristics	Float
5	**family_history_ with_overweight**	**Yes**: 1726 (81.7%) **No**: 385 (18.2%)	Characteristics	Object
6	**FAVC (**Frequent consumption of high caloric food)	Yes: 1866 (88.3%) No: 245 (11.6%)	Eating Habit	Object
7	**FCVC (**Frequency of consumption of vegetables)	Never: 102 (4.83%) Sometimes: 1013 (47.99%) Always: 996 (47.18%)	Eating Habit	Float
8	**NCP (**Number of main meals)	Once: 316 (14.97%) Twice: 176 (8.34%) Three times: 1470 (69.64%) Four times: 149 (7.06%)	Eating Habit	Float

9	**CAEC (**Consumption of food between meals)	Sometimes: 83.6% Frequently: 11.5% Always: 2.5% no: 2.4159	Eating Habit	Object
10	**SMOKE**	No: 2067 (98%) Yes: 44 (2%)	Whether Yes or Not	Object
11	**CH2O (**Consumption of water daily)	Less than a liter: 485 (22.97%) Between 1 and 2 L: 1110 (52.58%) More than 2 L: 516 (24.44%)	Eating Habit	Float
12	**SCC (**Calories consumption monitoring)	No: 2015 (95.4%) Yes: 96 (4.5%)	Physical Condition	Object
13	**FAF (**Physical activity frequency)	I do not have: 720 (34.11%) 1 or 2 days: 776 (36.76%) 2 or 4 days: 496 (23.50%) 4 or 5 days: 119 (5.64%)	Physical Condition	Float
14	**TUE (**Time using technology devices)	0-2 hours: 952 (45.10%) 3-5 hours: 915 (43.34%) More than 5 hours: 244 (11.56%)	Physical Condition	Float
15	**CALC (**Consumption of alcohol)	Sometimes: 1401 (66.3%) No: 30.2% Frequently: 3.3% Always: 0.047%	Eating Habit	Object
16	**MTRANS (**Transportation used)	Public Transportation: 1580 (74.8%) Automobile: 457 (21.6%) Walking: 56 (2.6%) Motorbike: 11 (0.5%) Bike: 7 (0.3%)	Physical Condition	Category

➢ *Correlation between Features*

This analysis seeks to establish how the feature variables relate to one another or correlate. The correlation coefficient is between −1, and 1. The closer the coefficient is to 1, the stronger the positive linear relationship between variables. The closer the coefficient is to −1, the stronger the negative linear relationship between variables. The closer the coefficient is to 0, the weaker the linear relationship between the variables.

The figure below illustrates the correlation between quantitative feature variables where one can notice that two variables 'Height and weight' are distinguished from others by having a stronger relationship or high correlation of 4.6 percent compared to others.

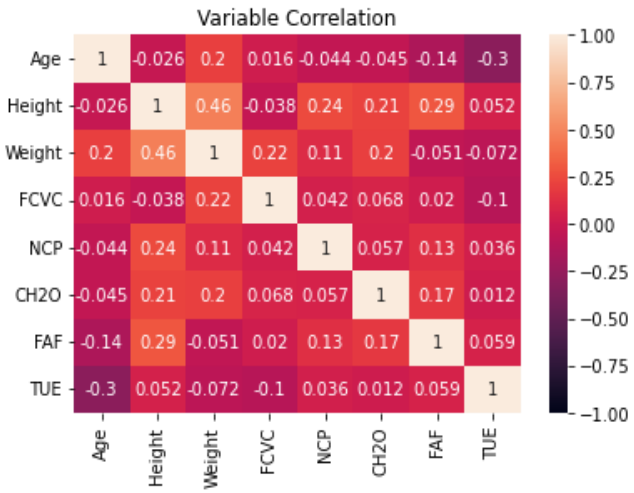

Fig 3 Correlation between Variables

Correlation coefficients are used to quantify the strength of an association between two variables [20].The relationship between two variables implies that when the value of one variable changes, another variable tends to simultaneously change. Understanding this connection is crucial because the value of one variable can be used to predict the value of the other. For instance, here, weight and height are correlated as height increases, so does weight. As a result, if a person is unusually tall (height), one can predict that they are also particularly heavy (weight) [21].

➢ *Linear Relationship Between Height and Weight*

A linear relationship (also known as a linear dependence) is a statistical phrase that refers to a straight connection between two variables. The linear correlation coefficient quantifies the strength and direction of a linear link between two variables [38][20]. This graphic below clearly confirms how the variables Height and Weight are in a linear relationship. This linear relationship is clear evidence that both variables are used to calculate the Body Mass Index (BMI) which is the other metric used to calculate obesity level within a population.

ISSN No:-2456-2165

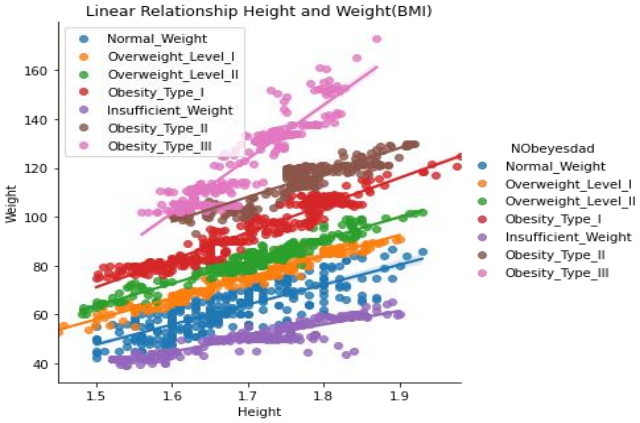

Fig 4 Linear Relationship between Variables Height and Weight (BMI)

The calculation of the weight-to-height-squared ratio is a reliable way to state whether a person has enough body fat. This ratio, known as the body mass index (BMI), leads to the conclusion that taller people have more tissues than shorter people, resulting in them weighing more [22].

➢ *Relationship Height-Weight with Target Variable (Boxplot)*

As shown in the figure below, we use a boxplot, which is a type of graph, to show how the values of the target variable are distributed in the dataset according to these two variables Height and Weight.

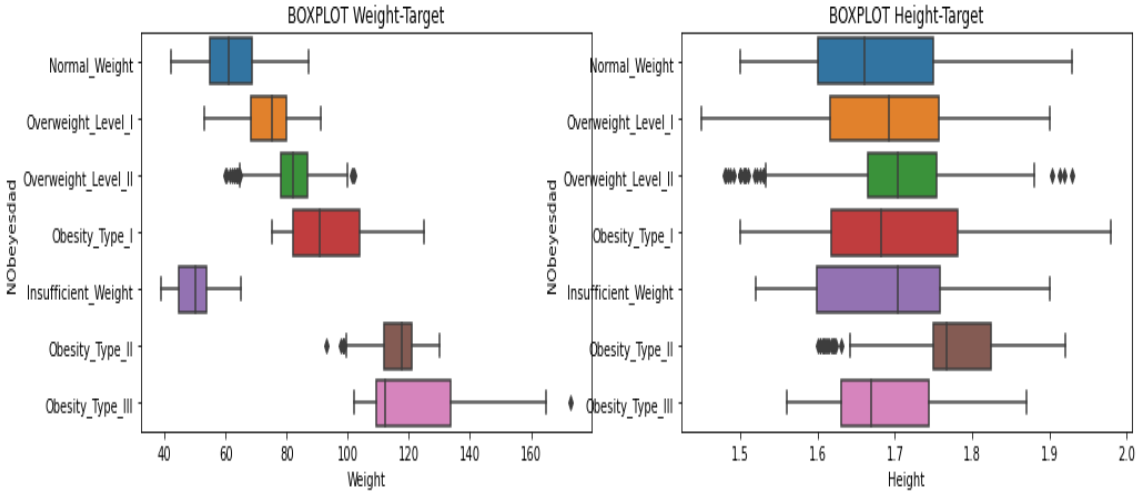

Fig 5 Boxplot Weight-Target and Boxplot Height-Target

This boxplot illustrates that many classes of the target variables are not normally distributed when we look at the position of the median. They are either in the negative or positive skew. The other important thing we notice through this boxplot is the presence of some data points which are out of the ordinary ones. These values are known as 'outliers' that we will seek to manage during the preprocessing and the modeling of the machine-learning algorithm because those outliers can affect the performance of some machine-learning algorithms. Outliers in a normal distribution, for example, could be values on the tails of the distribution [10].

E. Data Preprocessing

The quality of data has a substantial impact on the performance of the machine-learning algorithm. Thus, utterly worthless, incomplete, redundant, and similar data will undoubtedly harm their performance; this is why the data used must be preprocessed [25]. To create a dependable model with predictable performance, data preparation is one of the most crucial steps in the machine learning process. However, this procedure has the potential to either strengthen or weaken the prediction model [2]. This process implies encoding, Normalization, Imputation, feature selection, feature engineering, and other techniques.

In this part, we started by encoding which is the fact of converting categorical variables into numerical values so that they can then be easily fitted to machine-learning algorithms because machine-learning deals only with numeral data to make the computation. Other preprocessing processes were carried out interactively with the modeling process according to the requirements of every selected machine-learning algorithm.

➢ *Feature Importance*

Variable importance or feature importance analysis is one of the preprocessing methods that describe variables according to how relevant they can predict the target labeled variable [22]. This analysis facilitates a greater understanding of the dataset and, on the occasion, it conducts machine-learning model improvements during the feature selection process [22]. The illustration below shows the level of 'feature importance'. It is obtained during the preprocessing from the Random Forest Classifier. As one can see, the model gives the greatest importance to the variable Weight followed by Height, and Age of individuals. The lowest importance is given to variables SCC (Calories Consumption Monitoring), followed by FAVC (**(**Frequent consumption of high caloric food), **CALC (**Consumption of alcohol), SMOKE.

ISSN No:-2456-2165

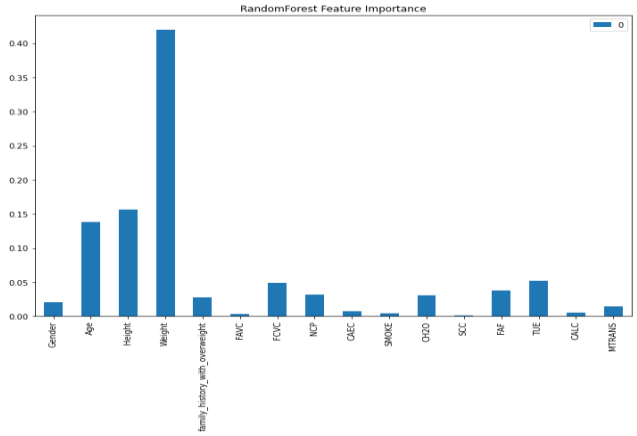

Fig 6 Random Forest Feature Importance

III. MODELING

Following the Exploratory Data Analysis (EDA) and the preprocessing, five machine-learning algorithms have been applied within a single list of models including K-Neighbors Classifier (KNN), Support Vector Machines (SVM), Ridge Classifier, Logistic Regression, and Random Forest Classifier.

To evaluate our list of models we used metrics as given by the Classification Report to well understand the overall performance of all the algorithms trained in one list of models, The confusion-matrix to be aware of 'true positive', 'true negative', 'false positive', and 'false negative' prediction. Finally, we used the learning curve to see the way the 'validation score' and the 'training score' converge throughout the learning process.

A. Model Validation

K-Fold Cross-Validation technique was used to separate the training dataset into four different (4-Fold Cross Validation) splits in which, inside each split one fold is considered as a validation set. To validate the performance of each model, K-fold calculates the mean average of each split and is considered the best model the one with the highest mean.

After having trained the five machine-learning algorithms with accuracy as the main metric we got these results: 91% of accuracy for Random Forest Classifier, 83% of accuracy for K-Nearest Neighbor Classifier, 83% of accuracy for Support Vector Machine, 80% of accuracy for Logistic Regression, and 67% of accuracy for Ridge Classifier.

Table 2 Model Validation

S No.	Machine-Learning Algorithms	Performance Accuracy
1	Random Forest Classifier	91%
2	K-Nearest Neighbors Classifier	83%
3	Support Vector Machine	80%
4	Logistic Regression	80%
5	Ridge Classifier	67%

➢ *Choice of the Final Model*

Mathematically speaking 91% is greater than 83% (Random Forest is greater than KNN). 80% is equal to 80% (Logistic Regression is equal to Support Vector Machine). And Ridge Classifier gives the lowest performance with 67% of accuracy.

Therefore, from the mathematical point of view according to their accuracy, the best performance accuracy is attributed to the Random Forest algorithm because of its highest performance compared to four other algorithms. However, accuracy is not the only factor to select an adequate algorithm. There are some other important criteria to observe to make an optimal decision of the machine learning algorithms i.e. the model that better suits the problem one needs to solve. Some of those criteria are the quantity of the available records in the dataset, the type of those data (qualitative or quantitative), the data normalization, Interpretability, Linearity of data, Training time, Data format, Memory requirements, The number of features, and Prediction time.

Talking about the criteria of linearity, during the Exploratory Data Analysis, it was observed a strong linear relationship between two variables "Height and Weight". These two variables again have been confirmed among the most important ones during the preprocessing while talking about "Feature Importance". In this situation of a linear relationship between variables, the support vector machine is one of the two best appropriate machine-learning algorithms; on the other hand Random Forest is among the inappropriate ones even if it can work sometimes. SVM uses a kernel trick to accommodate both non-linear and linear solutions [22].

Based on the above-mentioned criteria related to the choice of one model to optimize for the sake of our system and according to the objective of this study, we select the support vector machine (SVM) as the final machine learning algorithm of this research without denying that the Random Forest, the Nearest Neighbor, and even other machine learning models could solve this problem.

B. Final Model Optimization

The penalty factor C and the kernel parameters are referred to as hyper-parameters because they are at the top level of the SVM implementation; their main feature is that hyper-parameters describe the range of available decision functions[23]. We need here to deduce from this reliance to obtain the optimized hyper-parameters results in the best decision function for the Support Vector Machine model.

However, the issue of hyper-parameter tuning would be to find a combination of values that optimize some performance criteria, such as the accuracy index and the proportion of support vectors in the case of Support Vector Classification (SVC).

ISSN No:-2456-2165

C. Support Vector Machine Hyperparameter Tuning

For the selected model, two hyperparameters C and gamma have been optimized. Gamma is a hyperparameter used with nonlinear SVM. One of the most commonly used nonlinear kernels is the radial basis function (RBF), which is defined as a function that aims at assigning a real value to every single input from its domain (it is then a real-value function), and the value that RBF produced is always an absolute value; meaning that, it is a measure of distance and its value cannot be negative [24]. The RBF gamma parameter controls the distance from the influence of a single training point. Hypermeter C is used in SVM to control errors for a linear kernel. However, with the use of RBF kernel, C and gamma parameters have to be simultaneously optimized. If the gamma is large, the effect of C becomes negligible. If gamma is small, C affects the model as it affects a linear model. Typical values of C and gamma are as follows: 0.0001 <gamma <10 & 0.1 <C <100, however, specific optimal values may exist depending on the application.

Low gamma values indicate a large radius of similarity which results in the clustering of more points. For high values of gamma, the points should be very close to each other to be considered in the same group (or class). Therefore, models with very high gamma values tend to overfit.

D. Optimization Process

We imported from skit learn model selection GridSearchCV, which is a machine learning function used to find the best hyperparameters by comparing different performances for each combination thanks to the cross-validation method. After that, we created a dictionary of hyperparameters containing the penalty coefficient C, and gamma. From the key gamma and C, different values were tested. And then, we used Polynomial Features to generate variables (features) by the exponentiation of available features. Finally, we used the Feature Selection technique with the function Select K-best (k as the number of iterations) to pick up features that hold the greatest importance on the target variable.

IV. RESULTS AND DISCUSSION

After having evaluated our optimized Support Vector classifier in Grid Search CV with the same metrics as given by the Classification Report, the Confusion Matrix, and the learning curve, our model accuracy improved from 80% to 97% with the following f1-score as seen in the classification report below: 96% for class one (obesity type I), 99% for class two (obesity type II), 99% for class three (obesity type III), 95% for class four (overweight level I), 91% for class five (overweigh level II), 97% for class six (normal weight), and 99% for class seven (insufficient weight).

	precision	recall	f1-score	support
0	0.96	0.97	0.96	67
1	0.98	1.00	0.99	53
2	0.99	1.00	0.99	69
3	0.92	0.98	0.95	57
4	0.98	0.85	0.91	55
5	0.98	0.96	0.97	57
6	0.98	1.00	0.99	65
accuracy			0.97	423
macro avg	0.97	0.97	0.97	423
weighted avg	0.97	0.97	0.97	423

Fig 7 Optimized Support Vector Machine Classification Report

The optimized confusion matrix below corroborates the results given by the classification report as we can see from the first class of 67 data points representing individuals (patients), 65 good predictions and two errors; in the second class of 53 data points, 53 good predictions; the third class of 69 data points without error; in the fourth class 57 data points, 56 good predictions with only one error; in the fifth class of 55 data points, 47 good predictions with 8 errors; in the sixth class of 57 data points, 55 good predictions with 2 errors; and for the last class of 65 data points, 65 good predictions with 5 errors.

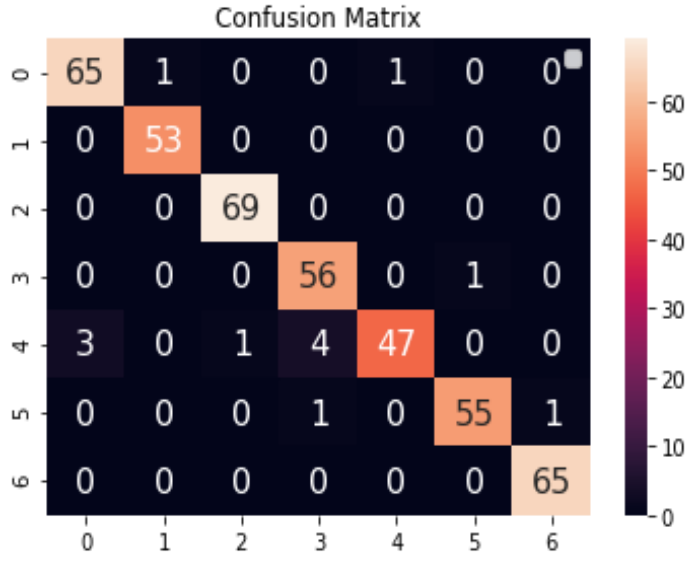

Fig 8 Optimized Support Vector Machine Confusion Matrix

ISSN No:-2456-2165

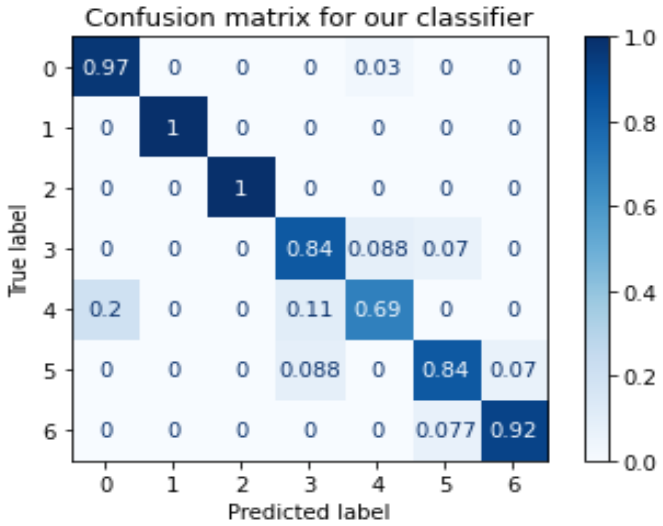

Fig 9 Optimized Confusion Matrix in percentage

The learning curve is completed while diagnosing the system after the classification report and the confusion matrix has shown similar situations. We can see that the system has a reasonable fit learning curve based on the dataset.

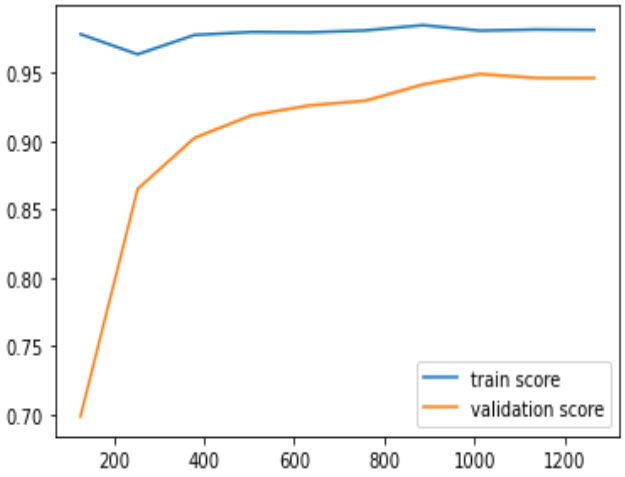

Fig 10 Optimized Support Vector Machine Learning Curve

V. CONCLUSION

Public health officials create education programs, make policy recommendations, manage services, and conduct research to prevent the occurrence or recurrence of health problems around the world. Other health professionals, such as doctors and nurses, prioritize patient care and prescription writing. Public health aims to reduce health disparities to improve healthcare quality, accessibility, and equity.

This research aimed at developing an ICT-based solution using machine-learning methodologies to support the public health policy for the early detection of obesity which literature confirmed as one of most world health threats causing chronic diseases and leading to premature deaths, especially in low-and-middle-income countries with inadequate health care system.

First and foremost from the selected dataset we carried out the exploratory data analysis (EDA) and data preprocessing to have an exhaustive understanding of different variables and convert data into machine learning format to facilitate the computation (learning). After that, five supervised machine learning classification algorithms were trained within a list of models including Ridge Classifier, Random Forest Classifier, K-nearest Neighbors, Logistic Regression, and Support Vector Machine to classify whether a person (patient) is obese, overweight, normal weight, or insufficient weight with same evaluation process while using metrics as given by the classification report, the confusion matrix, and the learning curve. Finally by comparing the performance of each of the models containing in the same list of models according to the results given by the classification report, the confusion matrix, and the learning curve; the support vector machine (SVM) was chosen for the best of our understanding as the final model that we optimized and received 97% of the performance accuracy.

Regarding low-income nations with inadequate health care coverage that are unlikely to provide broad access to essential metabolic illnesses that silently kill people, this ICT-based solution provided is particularly helpful. This ICT-based health management system will regulate metabolic disorders at a cheap cost and in real-time while sensibly reducing risk factors associated with the metabolic syndrome.

REFERENCES

[1]. Ritchie, H, and Roser, M (2022)Obesity. https://ourworldindata.org/obesity
[2]. WHO (9 June 2021) Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
[3]. Ritchie, H. (2022) What is obesity and how is it measured? www.OurWorldInData. Org/Obesity
[4]. WHO-Regional Office for Africa(2022)Obesity. https://www.afro.who.int/health-topics/obesity
[5]. WHO(13 April 2021)noncommunicable diseases. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
[6]. Ritchie, H, and Roser, M(2019)Causes of death. OurWorldInData
[7]. Paul Makan Mawaw et al.(2017)Prevalence of obesity, diabetes mellitus, hypertension and associated risk factors in a mining workforce, Democratic Republic of Congo. Pan African Medical Journal. doi: 10.11604/pamj.2017.28.282.14361
[8]. Halim N., Spielman K, Larson B. The economic consequences of selected maternal and early childhood nutrition interventions in low- and middle-income countries: a review of the literature, 2000—2013. BMC Women's Health, 2015, ISSN: 1472-6874.
[9]. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019 Jun;6(2):94-98. doi: 10.7861/futurehosp.6-2-94. PMID: 31363513; PMCID: PMC6616181.

[10]. Robert N. How artificial intelligence is changing nursing. Nurs Manage. 2019 Sep;50(9):30-39. doi: 10.1097/01.NUMA.0000578988.56622.21. PMID: 31425440; PMCID: PMC7597764.
[11]. Safaei M, Elankovan A. Sundararajan, Driss M, Boulila W, Shapi'i A. A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity, Computers in Biology and Medicine. j.combiomed. 2021, 104754, ISSN 0010-4825.
[12]. Ferdowsy F, Rahi KSA, Jabiullah I, Habib T. A machine learning approach for obesity risk prediction. Current Research in Behavioral Sciences, 2021, 100053, ISSN 2666-5182. https://doi.org/10.1016/j.crbeha.2021.100053.
[13]. Quiroz J.P.S, Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. Informatics in Medicine Unlocked, 2022, 100901, ISSN 2352-9148. https://doi.org/10.1016/j.imu.2022.100901.
[14]. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine Learning Techniques for Prediction of Early Childhood Obesity. Appl Clin Inform. 2015 Aug 12;6(3):506-20. doi: 10.4338/ACI-2015-03-RA-0036. PMID: 26448795; PMCID: PMC4586339.
[15]. Molina D, De-La-Hoz A, Mendoza F. Classification and Features Selection Method For Obesity Level Prediction. Journal of Theoretical and Applied Information Technology, 2021, ISSN: 1992-8645, E-ISSN: 1817-3195
[16]. Marcos-Pasero H, Colmenarejo G, Aguilar-Aguilar E, Ramírez de Molina A, Reglero G, Loria-Kohen V. Ranking of a wide multidomain set of predictor variables of children obesity by machine learning variable importance techniques. Sci Rep. 2021 Jan 21;11(1):1910. doi: 10.1038/s41598-021-81205-8. PMID: 33479310; PMCID: PMC7820584.
[17]. Palechor FM, Manotas AH. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru, and Mexico. Data in Brief, 2019,p. 104344, ISSN 2352-3409.
[18]. El-Hazmi MA & Warsy AS. Relationship between Age and the Prevalence of Obesity and Overweight in Saudi Population. Bahrain Medical Bulletin, Vol.24, No.2, June 2002
[19]. Jura M & Kozak LP. Obesity and related consequences to ageing. Epub 2016. PMID 26846415; PMC 5005878. doi: 10.1007/s11357-016-9884-3
[20]. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012 Sep;24(3):69-71. PMID: 23638278; PMCID: PMC3576830.
[21].]Frost J. Introduction to Statistics: An Intuitive Guide for Analyzing Data and Unlocking Discoveries.Epub, 2020, ISBN-13:978-1735431109
[22]. Hoyle B, Rau MM, Zitlau R, Seitz S, Weller J, Feature importance for machine learning redshifts applied to SDSS galaxies, Monthly Notices of the Royal Astronomical Society, Volume 449, Issue 2, 11 May 2015, Pages 1275–1283, https://doi.org/10.1093/mnras/stv373
[23]. Scikit-learn developers. "Validation curves: plotting scores to evaluate models — sci-kit-learn 0.20.2 documentation". Retrieved February 15, 2019.
[24]. Padierna LC, Carpio M, Rojas A, Puga H, Baltazar R, Fraire H.Hyper-Parameter Tuning for Support Vector Machines by Estimation of Distribution Algorithms. Nature-Inspired Design of Hybrid Intelligent Systems, 2017; ISBN:978-3-319-47053-5 DOI: 10.1007/978-3-319-47054-2_53
[25]. Bors AG. Introduction of the Radial Basis Function (RBF) Networks. Schwenker, Online Symposium for Electronics Engineers, 2001, 1(1):1-7.
[26]. Gupta, M et al (2019)Obesity Prediction with EHR Data: A deep learning approach with interpretable elements. ACM Transactions on Computing for Healthcare, 2022 https://doi.org/10.1145/3506719
[27]. Harvard T.H. Chan School of Public Health(2021).Obesity Prevention Source. Healthy Weight Report
[28]. Juneau, M(2019).Le transport actif, une excellente façon de concilier travail et santé
[29]. Kalra. V& Aggarwal.R. (2018).Importance of Text Data Preprocessing & Implementation in RapidMiner.ISSN 2300-5963 ACSIS, Vol. 14
[30]. Kim, E. et al. (2020) Application of Machine Learning to Predict Weight Loss in Overweight and Obese Patients on Korean Medicine Weight Management Program. Journal of Korean Medicine. DOI: https://doi.org/10.13048/jkm.20015