Artificial Intelligence

Title:- Estimation of Obesity Levels

M1 MIAGE 2IS 2025

Iana Miranda Caramé
iana.mica04@gmail.com
Hritik Bikram Rawal
hritik.rawal6642@gmail.com

Mintesnot Nigusu Yimer
kingyimer@gmail.com

Abstract

Obesity is one of the most serious problems affecting the world today. It's also a significant factor contributing to diabetes, heart disease, and other chronic ailments that are on the rise. To properly confront this problem, it is imperative to dive into the true causes of obesity and create robust classification models for its prediction with the help of machine learning tools.

The data set here has 2111 entries and has 17 distinct attributes. It has a range of demographic information like age and gender, physical information like height and weight, and behavioral variables like diet, physical activity, smoking, and alcohol intake.

The main focus of this work is to explore the success of different classification algorithms in the prediction of different categories of obesity. These results underscore the potential of machine learning for tracking the risk of obesity as well as for supporting diagnosis and intervention strategies at an earlier stage. Future work may consider the use of deep learning and inclusion of more behavioral or genetic parameters for better predictive performance.

Keywords: Obesity, Classification, PCA, Decision Tree, Random Forest, SVM, Logistic Regression

I.Introduction

Obesity is an alarming public health problem associated with various chronic illnesses requiring appropriate techniques for risk assessment and intervention. This investigation uses a set of 2,111 instances with 17 features composed of demographics (age, sex), physical features (height, weight), and behavioral features (diet, physical activity, smoking habits, and drinking) to classify seven levels of obesity. Since the target variable is categorical, this research applies supervised classification as opposed to regression; thus, the assignment of those cases is made to discrete obesity classes rather than predicting a continuous variable.

The models used for predictions include the Decision Trees, Random Forest, Logistic Regression and SVM, which are preferred due to their capabilities to process structured health data. Random Forest has a well-acknowledged ability to relieve the problems of overfitting and improve the model's generalizability. SVM is highly suited for complex and high-dimensional classification work. The above models are evaluated by metrics including accuracy, precision and recall to ascertain the most relevant variables.

This analysis is important since it applies machine learning for detecting key obesity determinants, rendering early-stage diagnosis, and informing health-related interventions according to data.

State of the art

The main goal of the recent research papers and analysis in this area was to classify and predict obesity levels by considering different factors like physical activity and nutritional habits. So these papers used different classification algorithms like random forest (RF) and extreme gradient boosting (XGBoost) to predict obesity levels because they can effectively handle the mix of categorical and numerical features present in the data, are easy to apply, have a high accuracy rate in many problems, and can be applied quickly. The previous studies also noted that associations were found between obesity levels and features. but for the scope of our project we chose decision tree, random forest and SVM because we are more familiar with these algorithms and they successfully helped us achieve our goal.

II.Data Analysis

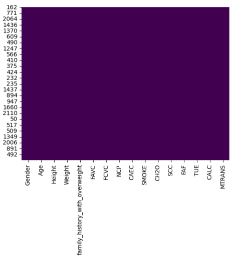
A. Features and labels

The dataset used in this paper has a collection of data estimating the obesity levels in people between the age 14 and 61 from Mexico, Peru and Columbia, based on their eating habits and physical conditions. The data contains 17 attributes and 2,111 records with the target NObeyesdad (Obesity Label) which allows to classify data based using the values Insufficient weight, Normal weight, Overweight Level I, Overweight Level II, Obesity type I, Obesity type II and Obesity type III. 23% of the data was gathered directly from users via a web platform, while 77% of the data was generated synthetically using the Weka tool and the SMOTE filter.

Name	**Role**	**Type**	**Description**	**Values**
Gender	Feature	Categorical	Gender	Male/Female
Age	Feature	Continuous	Contains Age	14 to 61
Height	Feature	Continuous	Height(meters)	1.45 – 1.98
Weight	Feature	Continuous	Weight(kg)	39 - 173
family_history_with_overweight	Feature	Binary	Has a family member suffered or suffers from being overweight?	Yes/No
FAVC	Feature	Binary	Do you eat high caloric food frequently?	Yes/No
FCVC	Feature	Integer	Do you usually eat vegetables in your meals?	Never, Sometimes, Always
NCP	Feature	Continuous	How many main meals do you have daily?	Between 1&2, 3, 3+
CAEC	Feature	Categorical	Do you eat any food between meals?	No, Sometimes, Frequently, Always
SMOKE	Feature	Binary	Do you smoke?	Yes or No

CH2O	Feature	Continuous	How much water do you drink daily?	<1L,1 to 2L, >2L
SCC	Feature	Binary	Do you monitor the calories you eat daily?	Yes or No
FAF	Feature	Continuous	How often do you have physical activity?	I don't, 1 or 2 days, 2 or 4 days, 4 or 5 days
TUE	Feature	Integer	How much time do you use technological devices such as cell phone, videogames, television, computer and others?	0-2h, 3-5h, More than 5h
CALC	Feature	Categorical	How often do you drink alcohol?	No, Sometimes, Frequently, Always
MTRANS	Feature	Categorical	Which transportation do you usually use?	Automobile, Bike, Motorbike, Walking,Public , Transport
NObeyesdad	Target	Categorical	Obesity level	Insufficient weight, Normal weight, Obesity type I, Obesity type II, Obesity type III, Overweight Level I, Overweight Level II

B. Missing Values Analysis



In this analysis, a thorough review of the dataset has been carried out to ascertain the presence of any missing values. After an exhaustive inspection of the information using a heatmap, as shown in the figure, we confirm that there are no missing values across the variables in this analysis. We can therefore proceed on an assumption that the data is consistent and fit to use for the next stage of a study.

C. Descriptive Statistics

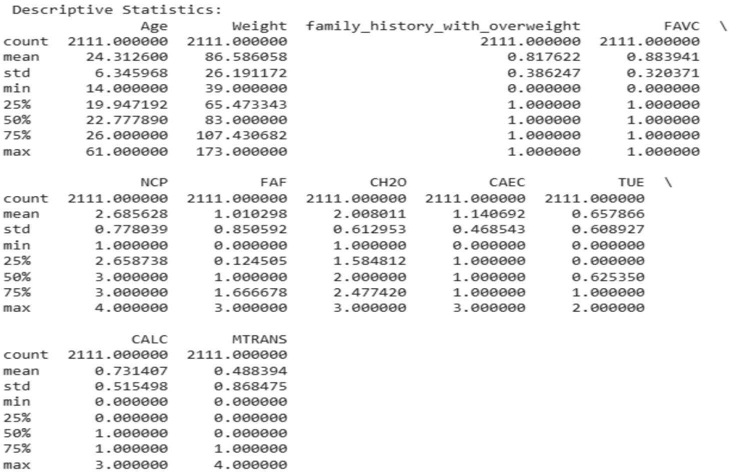
The dataset contains 2,111 observations with the 11 selected features consisting of both numerical and categorical variables related to obesity risk factors. The age of individuals ranges from 14 to 61 years with mean and standard deviation of 24.3 years and 8.42 years respectively. Weight ranges from 39 kg to 173 kg with the mean and standard deviation of 85.2 kg and 18.5 kg respectively.

The categorical features have been encoded numerically which includes family history of overweight, frequent consumption of high calorie foods (FAVC) and consumption of food between the meals (CAEC) categorized as never, sometimes, frequently, always has been encoded from 0 to 3. Similarly, Alcohol Consumption (CALC) follows the categorical distribution from 0 to 3.

Related to dietary habits, the Number of Main Meals per Day (NCP) has mean 2.68 meals/day and standard deviation of 0.778 while the Water Consumption (CH2O) ranges from 1.0L to 3.0L with a mean of 2.01L/day.

In terms of lifestyle, the physical activity frequency (FAF) varies from 0(no activity) to 3 (high activity), with the mean 1.01. The time spent using electronic devices (TUE) shows the mean 0.657 hours/day with some spending up to 2 hours per day. For Mode of Transportation (MTRANS), individuals primarily use public transport (0) or walking (1), with a smaller portion using automobiles, motorbikes, or bicycles (encoded 2-4).

Note: - (Gozukara Bag et al., 2023) Categorical data is converted into numerical values to ensure compatibility with ML models. It allows mathematical operations, improves model performance, and retains the ordinal nature of certain variables. Here, replace () is used instead of map () to avoid the return of 'NaN' values if there is any unexpected category



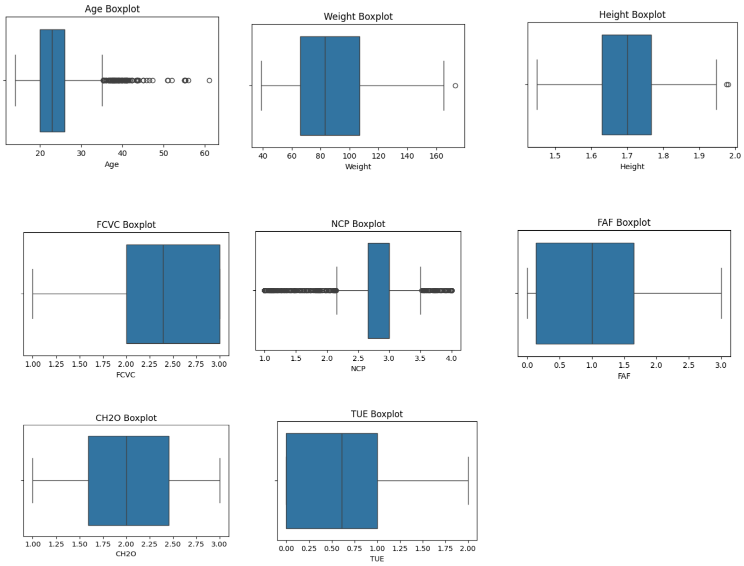
```
 Descriptive Statistics:
              Age       Weight  family_history_with_overweight         FAVC  \
count  2111.000000  2111.000000                     2111.000000  2111.000000
mean     24.312600    86.586058                        0.817622     0.883941
std       6.345968    26.191172                        0.386247     0.320371
min      14.000000    39.000000                        0.000000     0.000000
25%      19.947192    65.473343                        1.000000     1.000000
50%      22.777890    83.000000                        1.000000     1.000000
75%      26.000000   107.430682                        1.000000     1.000000
max      61.000000   173.000000                        1.000000     1.000000

              NCP          FAF         CH2O         CAEC          TUE  \
count  2111.000000  2111.000000  2111.000000  2111.000000  2111.000000
mean      2.685628     1.010298     2.008011     1.140692     0.657866
std       0.778039     0.850592     0.612953     0.468543     0.608927
min       1.000000     0.000000     1.000000     0.000000     0.000000
25%       2.658738     0.124505     1.584812     1.000000     0.000000
50%       3.000000     1.000000     2.000000     1.000000     0.625350
75%       3.000000     1.666678     2.477420     1.000000     1.000000
max       4.000000     3.000000     3.000000     3.000000     2.000000

             CALC       MTRANS
count  2111.000000  2111.000000
mean      0.731407     0.488394
std       0.515498     0.868475
min       0.000000     0.000000
25%       0.000000     0.000000
50%       1.000000     0.000000
75%       1.000000     1.000000
max       3.000000     4.000000
```

D. Outlier analysis

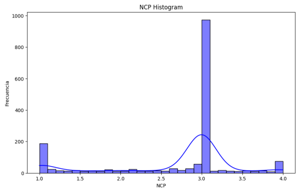
Outlier analysis is an important step in machine learning data preprocessing since outliers can affect the model's performances considerably. In this approach, we visually inspect the boxplots indicating the different outliers in the training dataset.



We start by isolating and selecting the numerical columns from the training dataset, which allows focus to be only on those variables where it is likely that outliers are present. We then generate boxplots that illustrate the median, interquartile range (75th and 25th percentiles), and the distribution of the data. Outliers are usually shown as isolated points outside of the whiskers on the boxplot. Here we see that TUE, FAF, CH2O, and FCVC possess no outliers while Height and Weight show very few. Particularly with these issues, they are probably extreme values within the features, yet they don't affect the overall distribution. However, with Age and NCP, the levels are significantly noticed with outliers.

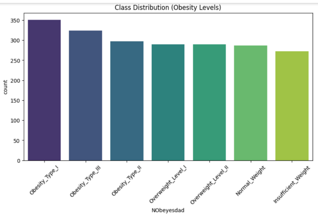
With respect to Age, the outliers remain within a sensible range, providing reasonable belief that they are due to the distribution just tending to the left. Because of this, no observations will be eliminated.

Nonetheless, we took it a step further as we examined the NCP feature in a little more detail and created a histogram for easy visualization. In the plot, one can see that the values range from a minimum of 1 to a maximum of 4; most of the values are clustered around 3. These all fall within an acceptable range for this feature, meaning any outliers that do show up are fairly legitimate and, certainly, not serious. So there is no need to remove these findings from the data set.



E. Balancing

There is a relatively small difference between the biggest number of examples for obesity Class 1 (350) and the smallest (290) for underweight. Many people consider a dataset unbalanced if the majority class outnumbers the minority class very significantly (think of proportions that go from 90:10 to as low as 80:20). Here, the proportions are 350/290 ≈ 1.21, implying that the two classes have quite similar sizes. Thus, from this perspective, the dataset is well balanced, and it is not necessary to institute any balancing method.



F. PCA

Principal Component Analysis (PCA) projects high-dimensional data onto fewer uncorrelated components, capturing directions of maximum variance. This helps visualize complex datasets in fewer dimensions and highlight broad patterns.

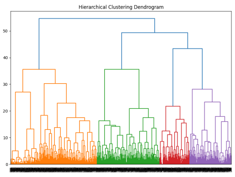
In the context of the obesity dataset, the PCA visualization shows how individuals with varying obesity distribute along the first two principal components. Some categories, such as Obesity_Type_III, appear relatively more separated in the PCA space, suggesting that certain underlying features (like weight and related measurements) contribute significantly to these principal components. Other classes, including Insufficient_Weight or Overweight_Level_I, tend to overlap, indicating that their differences may be driven by features not captured strongly by the top two components. This overlap can signal that additional principal components—or other analysis methods—may be required to distinguish these categories more clearly.

Since PCA is an unsupervised, linear method and therefore may not fully capture non-linear relationships or nuanced interactions among variables in the obesity dataset. Nonetheless, the PCA plot offers a useful starting point for understanding the global structure of the obesity dataset and for guiding subsequent, more targeted analyses.

G. Clustering

Hierarchical clustering seeks, through the agglomeration process, to analyze the natural structure of the data, performing no pre-tests to assume a number of groups. The algorithm works by iteratively combining similar observations, yielding a tree structure whose systematicity is seen, in the end, as a dendrogram. With the dendrogram, an intuition of the hierarchical relations among the data is drawn on the model obtained with scikit-learn's AgglomerativeClustering.

As the dendrogram was observed, evidence of distinct clusters was found suggesting possible cluster formations therein. This data helps determine the existence of an appropriate number of meaningful clusters behind the data to be further evaluated.

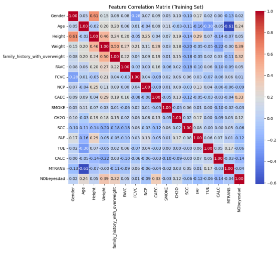


K-means clustering was also performed, but it behaved inferiorly in terms of interpretation due to its complete binary nature concerning many features. K-means computes based on distance, which can sometimes fail to replicate existing relationships in case of such distributions of data.

H. Correlation

Correlation helps in understanding the inter-relationship among the features within a dataset. Correlation coefficients quantify the strength and direction of the relationship between features. A correlation coefficient of 1 implies a strong positive correlation, while a correlation coefficient close to -1 illustrates a strong negative correlation. A correlation coefficient near zero denotes that the features do not possess any linear relationship.

We have generated a correlation matrix for all numerical features of the dataset. From the analysis represented in the figure, most of the correlations were not indicating that they were extremely strong, but nonetheless, some features may be related to one another to a limited extent.



As an example, weight expresses moderate correlation values with the family history of being overweight, which conforms to the idea that genetic predisposition may influence a person's body weight. In addition, features that relate more to lifestyle include frequency of caloric intake or frequency of physical activity, which express weak to moderate relationships with various health-related attributes. Conversely, exhibited limited correlation values mostly indicate that they are independent features to the dataset.

III. Machine Learning

A. Data leakage

Data leakage affects the model capability to generalize to new data. This implies that model performance is artificially raised due to training bias. Leakage may be caused by a number of reasons: this may be a representation mistake, contamination of the target variable, or improperly defined features. The several actions taken on our dataset ensure that our model is robust and generalizable for the application.

1. Splitting the data into test and training sets

Properly splitting the dataset into training and test sets before any other analysis is made is essential for avoiding data leakage. This avoids the transportation of any data from the test set back during model learning. With regards to preventing premature access of the test data by the training set, using test information too early could generate a falsely satisfying estimate on the performance of the model.

2. The Rate of Preprocessing

Before dealing with any arguments concerning data-leakage, one must first carry out the dataset into the training and test sets and apply any statistical transformations. Operations such as encoding categorical variables and feature scaling were applied only to the training data, and then the same transformations were transferred to the test set.

3. Categorical Feature Engineering

Afterward, either binary (like categorical variables such as gender, family history of overweight and/or transportation methods) or one-hot encoding were used. The binary encoding was applied to values that took only two values ("yes"/"no"), converted into values {0,1}; for categorical variables that have more than two unique values (for example, "should you consume calorie-dense foods" or "most favored transportation method"), each category was given a number uniquely.

As it was said before, to avoid data leakage, all encodings were performed only on the training set, and the same mappings were applied to the test set. This ensures that no information from the test data influences the training process, maintaining the model’s generalization ability.

B. Target justification

The dataset includes information on various obesity types, collected from a diverse group of individuals, with the goal of developing predictive models based on different algorithms. By analyzing features such as age, gender, lifestyle factors, and family members with a history of obesity, we aim to accurately predict the type of obesity present in patients. In the field of healthcare, the ability to predict obesity types can be immensely beneficial for clinicians and healthcare providers. It can aid in personalizing treatment plans, improving patient outcomes, and enhancing preventive strategies. Additionally, automated predictions can streamline the diagnostic process, allowing healthcare professionals to allocate their time and resources more efficiently. By ensuring timely and accurate predictions, we can reduce the risk of missing crucial patient information that is vital for effective treatment and intervention, ultimately contributing to better healthcare delivery and patient well-being.

C. Applied algorithms

In the following paragraphs it will be explained in detail which three algorithms were tested and applied during the course of this project: Support Vector Machine (SVM), Decision Tree/Random Forest, and Logistic Regression.

1. Support Vector Machine (SVM)

The Support Vector Machine is a supervised machine learning algorithm that aims to find an optimal hyperplane that best divides the classes by maximizing the margin. This actually serves to separate the classes of obesity with a great amount of ease. Unlike simpler models, SVM has shown a capacity to handle non-linear relationships amongst feature variables using kernel functions such as RBF Kernel. This piece of information is especially beneficial when a dataset entails numerous interacting factors since obesity is affected by diet, exercise, family history, among others.

Some disadvantages of SVM are its computationally expensive feature, which is mostly noticeable for large data sizes, especially when kernel tricks are employed, something which can suffer from time delays. SVM models are not as interpretable compared to other models such as Decision Trees. This makes it difficult for the user to draw conclusions from the outputs of the model.

Within these limitations, SVM is still an appropriate algorithm for this dataset due to its ability to capture complex patterns yet with a low risk of overfitting.

2. Decision Tree

A decision tree is a supervised learning algorithm that is simple yet effective, which makes predictions by splitting the data using features organized by importance sequentially into a tree of decision rules. This is quite interpretable, computationally efficient, and fits both numeric and categorical features, which supports its selection for obesity classification.

In applications such as medical diagnosis or obesity classification, decision trees allow practitioners to trace the decision-making process, offering transparency that is often required in clinical settings. They require minimal data preparation (for instance, no need for feature scaling). However, while they are fast and easy to implement, single decision trees can overfit the training data if not pruned properly, which is a significant trade-off compared to more robust models like Random Forests.

Random forests, too, pose some difficulties, as training multiple decision trees requires greatly increased computational power compared to a single tree; hence, this approach is computationally heavy.

Even though Random Forest is slightly more accurate than Decision Tree with 95.51% accuracy, we chose Decision Tree because it serves as a strong baseline due to their straightforward implementation and clarity for this project.

3. Logistic Regression

Logistic regression is a simple, yet powerful, classifier since it models the class membership probability with the logistic (sigmoid) function. It is extremely useful for binary classifiers but it can be extended into multiclass classification through techniques, such as one-vs-rest (OvR) or softmax regression. Since obesity classification uses ordered categories, logistic regression is providing a baseline model for interpretability and performance benchmarks.

Logistic regression is computationally easy and produces coefficients that convey the importance of features, thus easy to interpret. This enables the analysis of how different lifestyle factors (like, for example, calorie intake or physical activity) affect obesity levels.

On the other hand, logistic regression assumes a linear relationship between the predictors and the log odds of the target variable which can limit its capacity to capture some sophisticated patterns. This makes it less accurate than other models such as support vector machines and random forests that tend to perform better when relationships between variables are more complicated.

However, due to its simplicity, efficiency, and interpretability, logistic regression acts as baseline modeling in obesity classification.

D. Metrics

In order to evaluate our algorithm performance, we used the following metrics for obesity types: accuracy, precision, recall, and the confusion matrix.

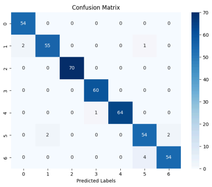
While accuracy gives a general measure of a model performance by relating the number of correct predictions to the number of actual predictions made, precision is the actual number of true positive observations to the number of actual ones predicted. That indicates how many of the obesity diagnosis predicted actually turned out to be true. Recall is an indicator of how well the model could find true positives among all actual obesity cases, demonstrating its sensitivity towards finding correct diagnoses. A high score in recall is crucial since failure to identify the correct type of obesity could result in subjective treatment which might impact the patients health. The confusion matrix allows a detailed report for the model’s predictions in terms of counts of true positives, true negatives, false positives, and false negatives. This breakdown is useful for understanding where the model's misprediction may arise, considering it also supports this analysis by visually displaying the perfect scenario for model evaluation and supplying the needed information for the computation of precision, recall, and accuracy.

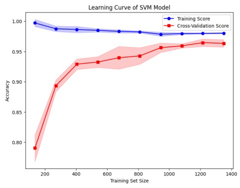
E. Training and Optimization

1.SVM

To optimize the SVM model, the regularization parameter C ranged from 0.1 to 10 while different kernel types like linear, RBF, polynomial were explored, and the gamma values used were either scale or auto. In order to perform the optimization we used GridSearch, a technique for tuning hyperparameters based on cross-validation to find the best combination for a model.

The optimizations lead to the conclusion that the model will perform well with this set of parameters: linear kernel, C=10, and gamma=scale which gave a cross-validation accuracy score of 0.9597. The SVM model demonstrates a very good performance in classifying the seven classes of the "NObeyesdad" target variable. The average overall model accuracy is 96%, with average precision and recall values also around 96% or above for all classes. Hence the model correctly identified most instances while a few were false positives or false negatives, as supported by the confusion matrix where most of the prediction results lay along the diagonal that represented right classification.





To assess whether the SVM has performed overfitting or underfitting, the training accuracy against test accuracy was examined; a big gap between the training and test accuracy would imply that the SVM is overfitted. In contrast, a low training and test accuracy will lead to underfitting. This would mean that a model fails to capture the underlying patterns because it is too simple. However, the model is well-generalized, makes very few misclassifications, and shows a consistent performance across all classes, even those with lower instances.

A learning curve was also plotted to visualize how model performance changes as the amount of training data increases. It is again shown that the model is well-generalized and that there’s no under or overfitting.

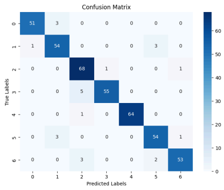
2. Decision tree

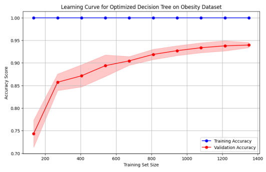
The decision tree classifier was initially trained without hyperparameter tuning, and it achieved an overall accuracy of approximately 90.31%. While this performance was promising, further improvements were desired, so we decided to optimize the model using RandomizedSearchCV with a reduced parameter grid. The inclusion of ccp_alpha parameter is particularly significant, as it controls the pruning of branches that provide little predictive power, thereby preventing overfitting.

The optimized model achieved a test accuracy of approximately 94.32%, with notable improvements across precision and recall for all obesity classes.

The best parameters were chosen because they struck an optimal balance between model complexity and generalizability. For example, the selected maximum depth prevented the tree from becoming too deep, reducing the risk of overfitting, while the tuned values for min_samples_split and min_samples_leaf ensured that each decision node was statistically robust. Moreover, the adoption of the ccp_alpha parameter allowed for effective cost-complexity pruning, which eliminated unnecessary branches and improved model performance on unseen data.

The confusion matrix reveals high overall accuracy, with most predictions correctly placed along the diagonal, indicating that the model effectively captures the underlying patterns in the obesity data. Misclassifications are minimal and primarily occur between adjacent weight classes. Notably, Obesity_Type_III has a precision of 1, underscoring its highly distinctive characteristics. Overall, while performance is strong, the few misclassifications suggest that further feature engineering could enhance the model's ability to distinguish closely related classes.



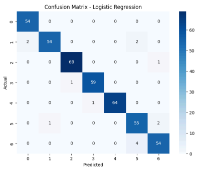


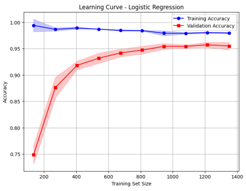
The learning curves show that the optimized decision tree achieves nearly perfect training accuracy across all training sizes indicating that although the model fits the training data very well, the tuning of hyperparameters effectively controls overfitting, allowing the model to capture complex patterns while generalizing well on unseen data.

3. Logistic Regression

Firstly, there was an attempt without optimization which resulted in an 86.76% accuracy. We then decided to optimize the Logistic Regression model, testing different values for the regularization parameter C (0.01 to 100) and solvers (liblinear, lbfgs) using GridSearch with cross-validation. The best parameters found were C=100 and solver=lbfgs, achieving a cross-validation accuracy of 96.69%.

The model performed well in classifying the seven categories of the "NObeyesdad" target variable, with an overall accuracy of 96.69% and both precision and recall above 96% for most classes. The confusion matrix showed that most predictions were correct, with few misclassifications.





To check for overfitting or underfitting, training and test accuracy were compared. The results indicated good generalization, as both accuracies were close. Additionally, the learning curve confirmed that the model learned effectively without significant overfitting or underfitting.

F. Results

All three models performed well, with accuracies above 90%. This indicates that the dataset does contain important patterns that can be adequately captured. The small differences in performance indicate that some feature interactions are better addressed by specific models. That High accuracy of Logistic Regression and SVM suggests that the dataset could consist of well-separated decision boundaries, thus suited for these algorithms. In contrast, for medical applications, it is essential to achieve a trade-off between accuracy, interpretability, and computational efficiency that might favor DT and LR over SVM, as SVM is not as interpretable and more computationally expensive.

These considerations make either Logistic Regression or Decision tree relatively good options, presenting very high accuracy and interpretability with valuable application in Medicine. However, should accuracy need to be serviced, LR might be the best choice, with the caution that it cannot assume anything other than a linear relationship and may therefore be outdone by a more capable model in handling complex decision boundaries.

IV. Conclusion

The study successfully applied machine learning techniques to predict obesity levels with high accuracy. The dataset contained valuable health and lifestyle information that contributed to model performance. The goal of using machine learning in this context is to provide an automated and efficient way to assess obesity risk, which is crucial for early intervention and public health strategies.

The use of SVM, Decision Tree, and Logistic Regression is well-founded since they offer distinctive advantages in accuracy, interpretability, and computational efficiency. While all models perform adequately, the best results in terms of accuracy were provided by SVM and Logistic Regression. Future work can explore other features, ensemble methods, or deep learning approaches to improve even further the performance and predictive strength of the models.

References

Gozukara Bag, H. G., Yağın, F. H., Gormez, Y., González, P. P., Çolak, C., Gülü, M., Badicu, G., & Ardigò, L. P. (2023). *Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits*. MDPI.

Quiroz, J. P. S. (2022). Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. *ScienceDirect*, 9. www.elsevier.com/locate/imu