University of Milano
Data Science for Economics
Statistical Learning Project

Selen Gezginci
32204A

Predicting Obesity Levels: A Study of Supervised and Unsupervised Machine Learning Techniques Using Eating Habits and Physical Condition Data

Abstract

This study explores the application of supervised and unsupervised machine learning techniques to predict and understand obesity levels based on individuals' eating habits and physical condition. Various machine learning algorithms were used, including Logistic Regression (Multinomial), K-Nearest Neighbors (KNN), LightGBM and Decision Tree, using the Target variable Obesity Level and a dataset consisting of 17 attributes and 2111 records. Additionally, K-Means clustering was utilized as an unsupervised learning method to identify distinct groups of individuals based on similarities in their characteristics. The study's findings reveal that logistic regression exhibits the highest accuracy, precision, recall, and F1 score among the supervised learning methods, indicating its effectiveness in predicting obesity levels. Moreover, K-Means clustering identified four distinct clusters within the dataset, shedding light on different features segments based on their attributes. These findings contribute to a better understanding of obesity and offer insights for tailored prevention and intervention strategies.

Keywords: Obesity, prediction, supervised, unsupervised, statistic, analysis

1. Introduction

Obesity is a complex condition characterized by excessive fat accumulation that can negatively impact overall health. Body Mass Index (BMI) is a commonly used measurement to assess overweight and obesity; A BMI of 25 or higher is considered overweight, and a BMI of 30 or higher is considered obese. Advanced analytical techniques, especially machine learning, can play an important role in predicting and managing obesity and thus contribute to more effective prevention and treatment strategies. Therefore, advances in obesity management can play an important role in improving both individual and societal health.
Machine learning algorithms are designed to extract information from data, and numerous research papers have confirmed their effectiveness in predicting and managing obesity. Machine learning can also be used to predict future outcomes, such as the likelihood of obesity-related health problems, and to help healthcare professionals design more effective interventions.

Overall, machine learning has the potential to significantly improve our understanding of obesity and contribute to the development of more effective prevention and treatment strategies. The aim of this study is to create a predictive model using machine learning techniques for the rapid detection and evaluation of overweight or obese individuals. Integrating machine learning into healthcare can improve diagnostic and treatment outcomes and provide a more effective approach to analysing and interpreting large amounts of data.

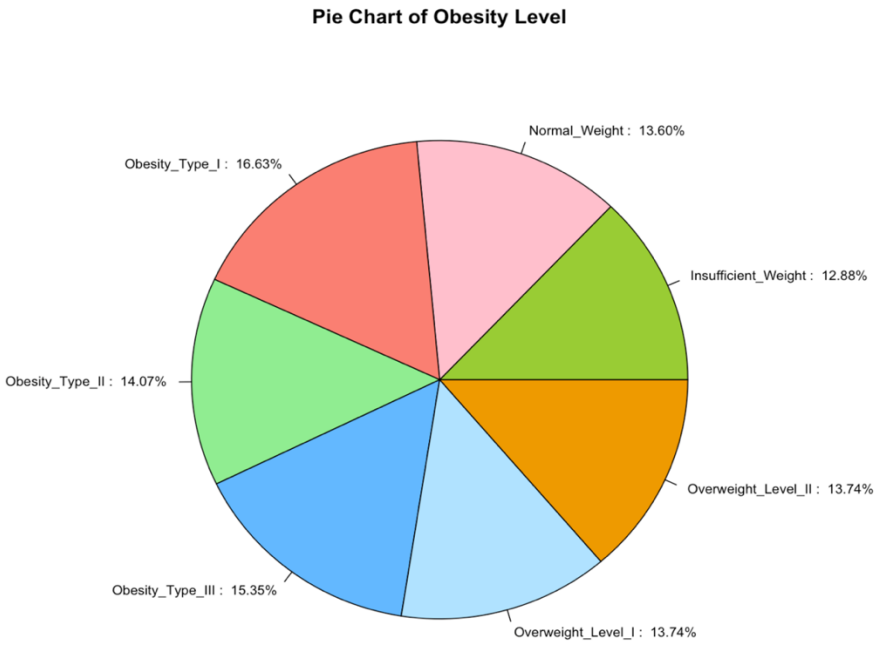
2. Data Information

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

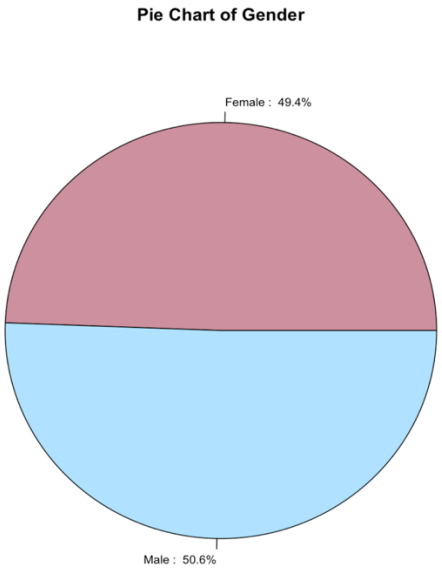
This dataset include data for the estimation of obesity levels in individuals, based on their eating habits and physical condition. The data contains 17 attributes and 2111 records, the records are labeled with the class variable NObeyesdad (Obesity Level), that allows classification of the data using the values of Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Insufficient_Weight, Obesity_Type_II, Obesity_Type_III. A single target provides data on a person's level of obesity. This dataset was used to train and evaluate the machine learning models applied in this study to predict the level of obesity in individuals. Each model input integrates a unique set of patient information along with a specific target variable. Input variables used in the model include Age, Gender, Physical activity, Consumption behaviours, Time using technology devices and Transportation used.

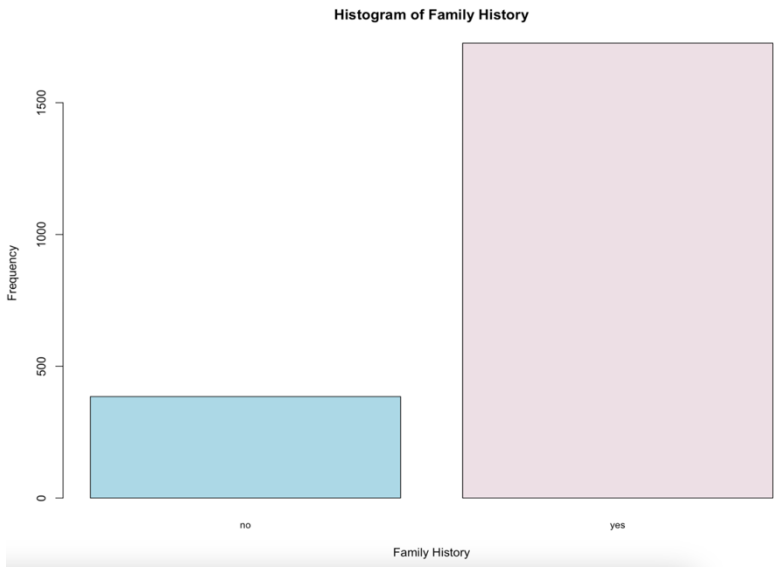
Variable	Data Type	Description	Values
Gender	character		Female, Male
Age	numeric		
Height	numeric		
Weight	numeric		
family_history_with_overweight	character		No,Yes
FAVC	character	Frequent consumption of high caloric food	No,Yes
FCVC	numeric	Frequency of consumption of vegetables	
NCP	numeric	Number of main meals	
CAEC	character	Consumption of food between meals	No, Sometimes, Frequently, Always
SMOKE	character		No,Yes
CH2O	numeric	Consumption of water daily	
SCC	character	Calories consumption monitoring	No,Yes
FAF	numeric	Physical activity frequency	
TUE	numeric	Time using technology devices	
CALC	character	Consumption of alcohol	No, Sometimes, Frequently, Always
MTRANS	character	Transportation used	Public_Transportation, Walking, Automobile, Motorbike, Bike
NObeyesdad	character		Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Insufficient_Weight, Obesity_Type_II, Obesity_Type_III

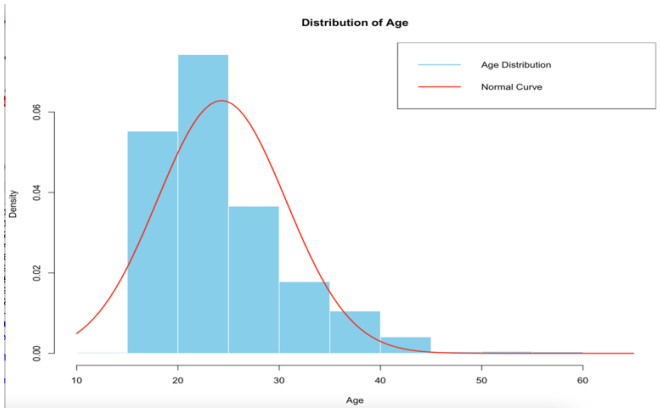
3. Data Explotary Analysis



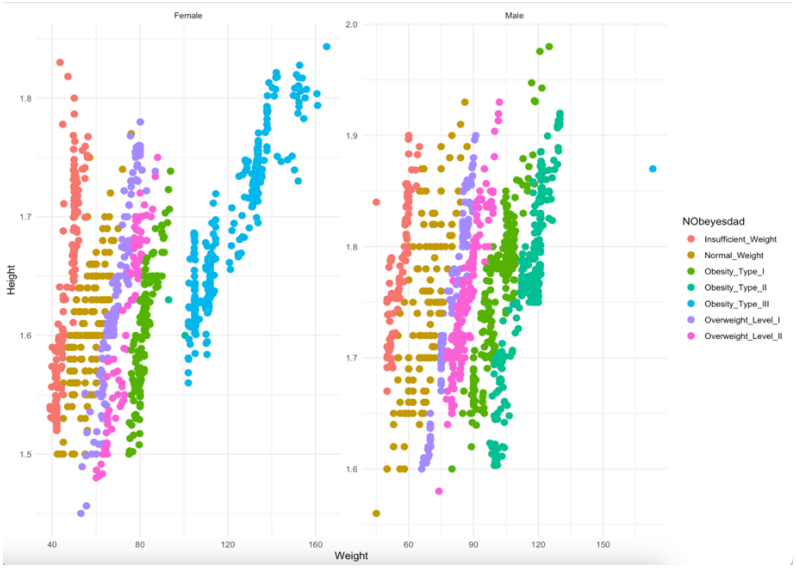
The pie chart shows that the distribution of obesity level within a dataset, with the majority falling into the obesity categories, comprising 46.05% in total (Type I: 16.63%, Type II: 14.07%, Type III: 15.35%). Additionally, there are segments representing overweight individuals, accounting for 27.08% (Level I: 13.74%, Level II: 13.74%), and those classified as having insufficient weight, making up 12.88% of the records.







The graph illustrates that The Distribution of Age and Age Distribution in the dataset.The most common ages are10 and 20, with a density of 0.06. The age 40 has lower density, but they are still relatively common compared to the ages between 10 and 40.



Gender	NObeyesdad	n
Female	Obesity_Type_III	323
Male	Obesity_Type_II	295
Male	Obesity_Type_I	195
Male	Overweight_Level_II	187
Female	Insufficient_Weight	173
Female	Obesity_Type_I	156
Male	Normal_Weight	146
Female	Overweight_Level_I	145
Male	Overweight_Level_I	145
Female	Normal_Weight	141
Female	Overweight_Level_II	103
Male	Insufficient_Weight	99
Female	Obesity_Type_II	2
Male	Obesity_Type_III	1

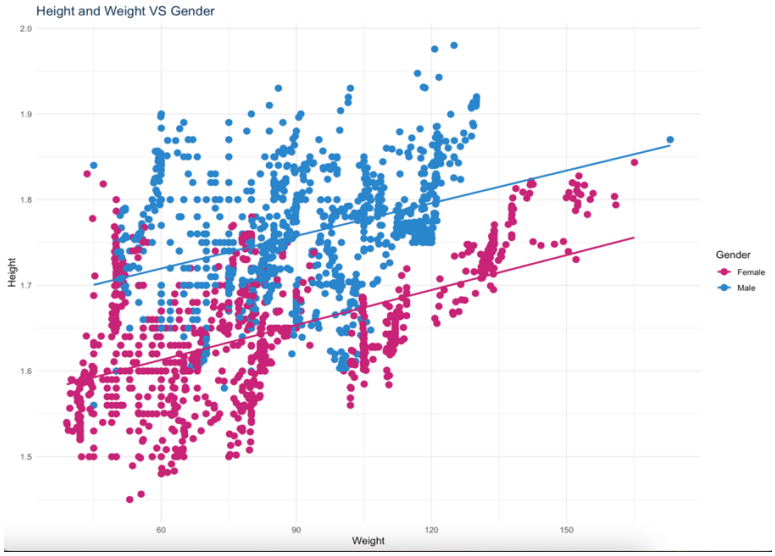
There are two the plot, one for women and the other for men.

Most data points for females are between 1.5 m and 1.8 m tall and weigh between 40 kg and 80 kg. This zone corresponds to the "Normal Weight" classification.Fewer data points for women are in the “Underweight” and “Overweight_Level_I” categories.There are only a few

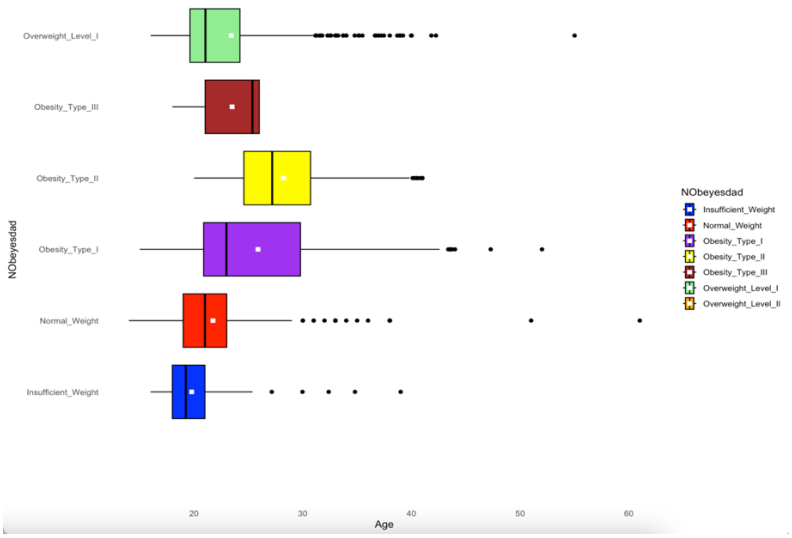
data points for women in the "Obesity_Type_II" and "Obesity_Type_III" categories and no data points in the "Obesity_Type_I" category.

Most data points for men are between 1.6 m and 1.9 m tall and weigh between 80 kg and 150 kg. This region corresponds to the "Overweight_Level_I" and "Obesity_Type_I" categories. Fewer data points for men are in the “Normal Weight” and “Obesity_Type_II” categories. There are no data points for men in the "Underweight" and "Obesity_Type_III" categories.

We can say that males tends to weigh more than women for any given height.



The scatter plot shows that the relationship between height, weight, and gender. The first column shows the height in meters, ranging from 1.5 to 2.0.The second column shows the weight in kilograms, with two values listed for each height: pink for females and blue for males. The third column shows the gender, with two values listed: "Female" and "Male". We can say that males tend to weigh more than females for a given height.



The box plot displays that the distribution of a age and obesity level.The horizontal axis represents the Obesity Level classifications. The vertical axis represents the age values, ranging from 20 to 60.

The median age for the Insufficient Weight category is lower than the median age for the Normal Weight category, which is lower than the median age for the Overweight_Level_I category. The median age continues to increase for the Overweight_Level_II and Obesity_Type_I categories, but then decreases slightly for the Obesity_Type_II and Obesity_Type_III categories.The IQR for the Insufficient Weight category is relatively small, indicating that the ages in this category are more consistent with each other. The IQR increases for the Normal Weight and Overweight_Level_I categories, indicating more variability in age. The IQR decreases slightly for the Overweight_Level_II and Obesity_Type_I categories, but then increases again for the Obesity_Type_II and Obesity_Type_III categories.

4. Data Preprocessing

There is no missing value in the dataset.

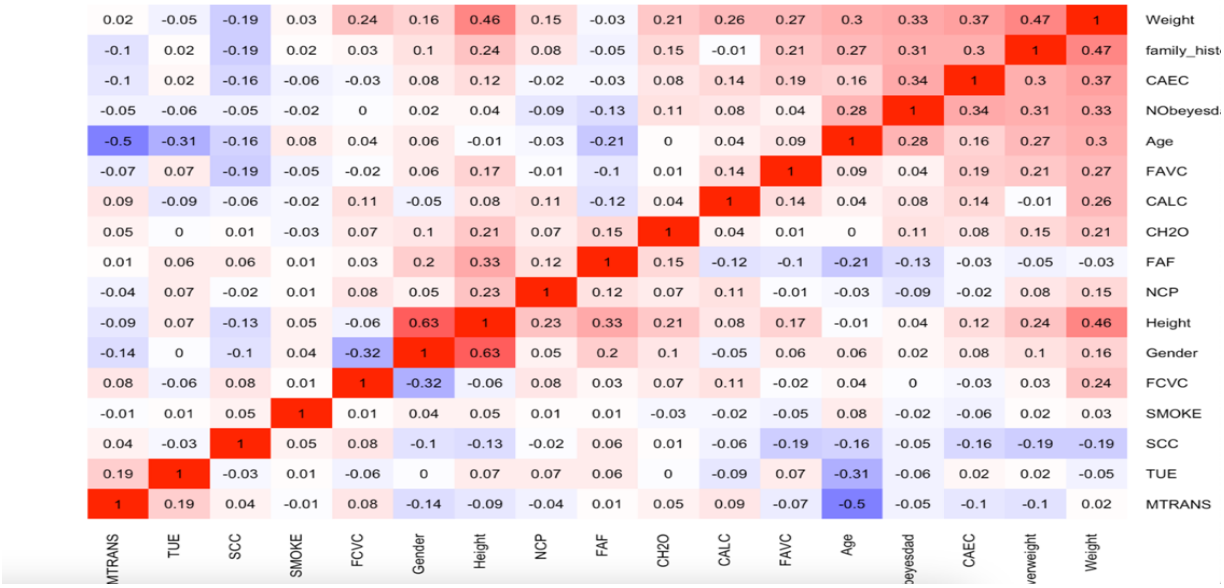
Categorical variables were encoded. Categorical variables were converted into numerical format using Yeo Johnson encoding technique.

The dataset was scaled to prevent one feature from dominating others during model training.

Correlation Matrix

Age, Consumption of food between meals, Physical activity frequency has negatively correlated with Obesity Level, suggesting that as these variables increase, Obesity Level tends to decrease.

Age, Consumption of food between meals, family history, Weight, and Consumption of water daily , as they seem to have a significant impact on Obesity Level. Family history, Weight, and Consumption of water daily have moderate positive correlations with Obesity Level. these variables increases, Obesity Level also tends to increase.



5. Machine Learning Algorithms

5.1. Supervised Learning Methods

5.1.1. Logistic Regression

In this study, logistic regression helps understand how different factors like eating habits and physical condition contribute to seven obesity levels. With a dataset of 17 attributes and 2111 records, LR offers computational efficiency. Using multinomial logistic regression, it can predict and classify individuals into seven obesity level categories such as Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Insufficient_Weight, Obesity_Type_II, Obesity_Type_III. This approach allows for a more detailed understanding of the factors affecting different levels of obesity, facilitating targeted prevention and intervention strategies specific to each category

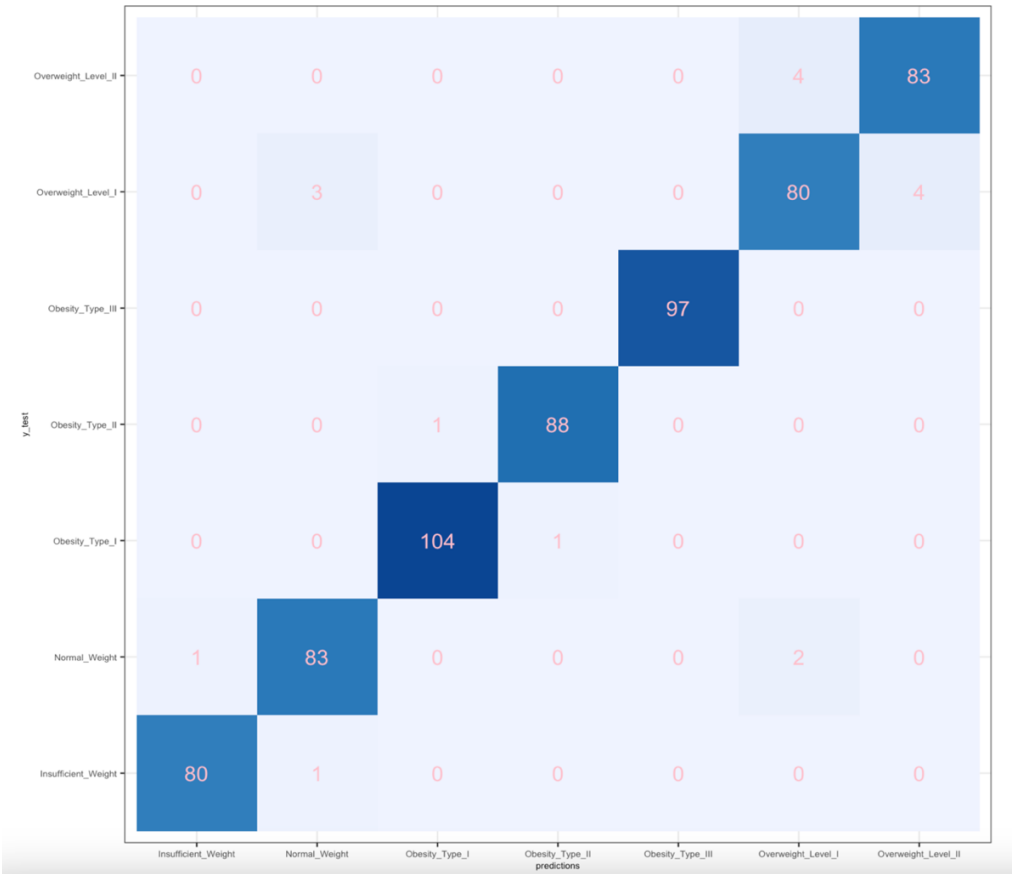
Accuracy: 0.9731013

Precision: 0.9721676

Recall: 0.9722249

F1 Score: 0.9721872

Confusion Matrix for logistic regression



5.1.2. KNN

K-Nearest Neighbors (KNN) is a simple yet powerful machine learning algorithm used for classification and regression tasks.KNN classifies data points by comparing them to their nearest neighbors. In this research, KNN helps identify patterns among individuals with similar characteristics, supporting in understanding factors contributing to different obesity levels. With its flexibility and efficiency, KNN complements traditional methods, offering insights for targeted prevention and intervention strategies.

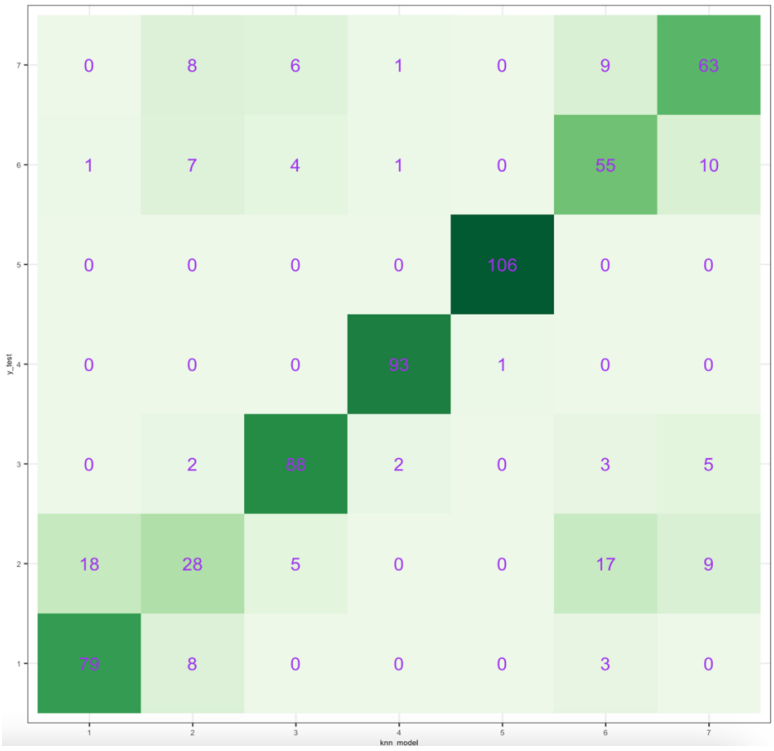
Accuracy: 0.8132911

Precision: 0.8026029

Recall: 0.8086387

F1 Score: 0.8020506

Confusion Matrix for KNN



5.1.3. LGBM

LightGBM builds decision trees vertically, meaning it grows leaf-wise instead of level-wise, which reduces the loss function more effectively and results in faster training times. LightGBM offers high accuracy and flexibility, making it suitable for a wide range of machine learning tasks, including classification, regression, and ranking.

In this study, LightGBM was used to analyze the data set with 17 features. This makes it a valuable tool for uncovering patterns and relationships between variables that contribute to obesity levels.

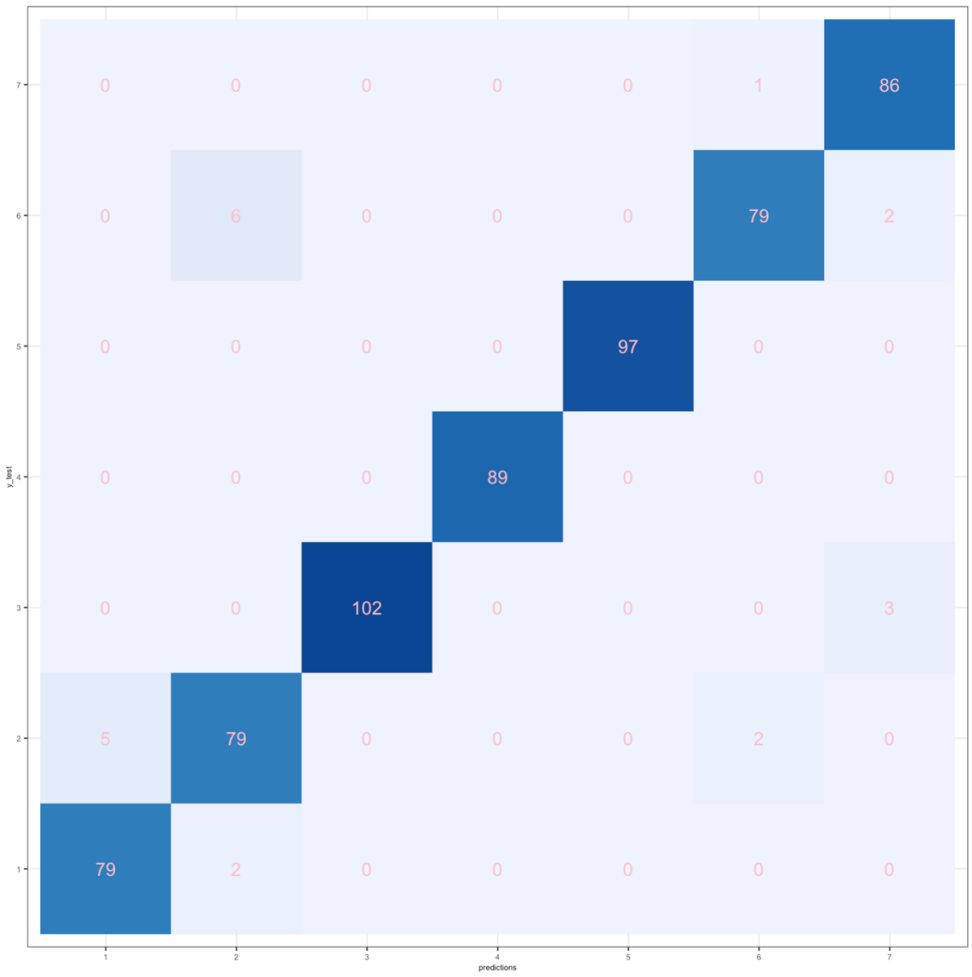
Accuracy: 0.9667722

Precision: 0.9652845

Recall: 0.9659848

F1 Score: 0.9653687

Confusion Matrix for LGBM



5.1.4. Decision Tree

Decision Tree is a machine learning algorithm used for both classification and regression tasks. It's a tree-like model where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents the outcome or prediction.

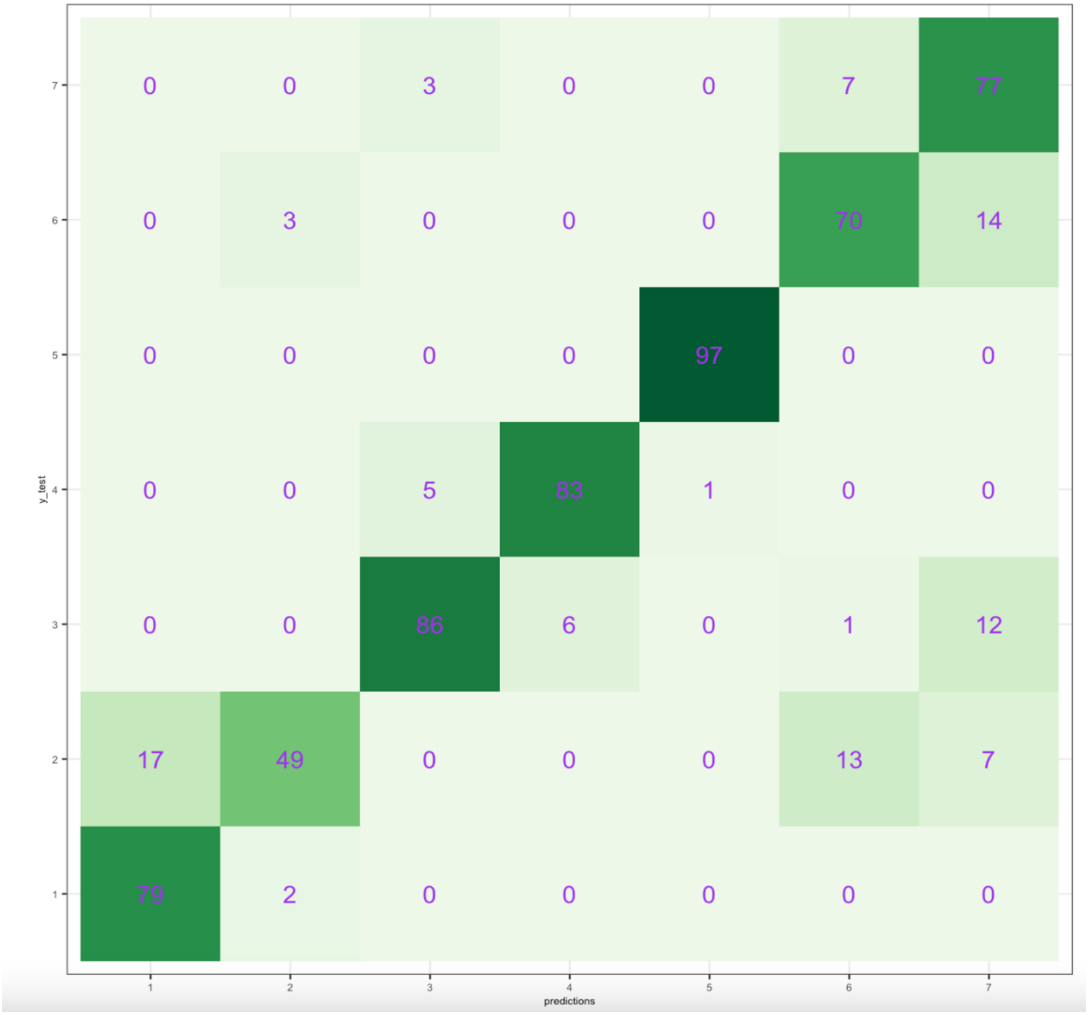
Accuracy: 0.8560127

Precision: 0.8624041

Recall: 0.8551947

F1 Score: 0.8503823

Confusion Matrix for Decision Tree



5.1.5. Comparision of Supervised Methods

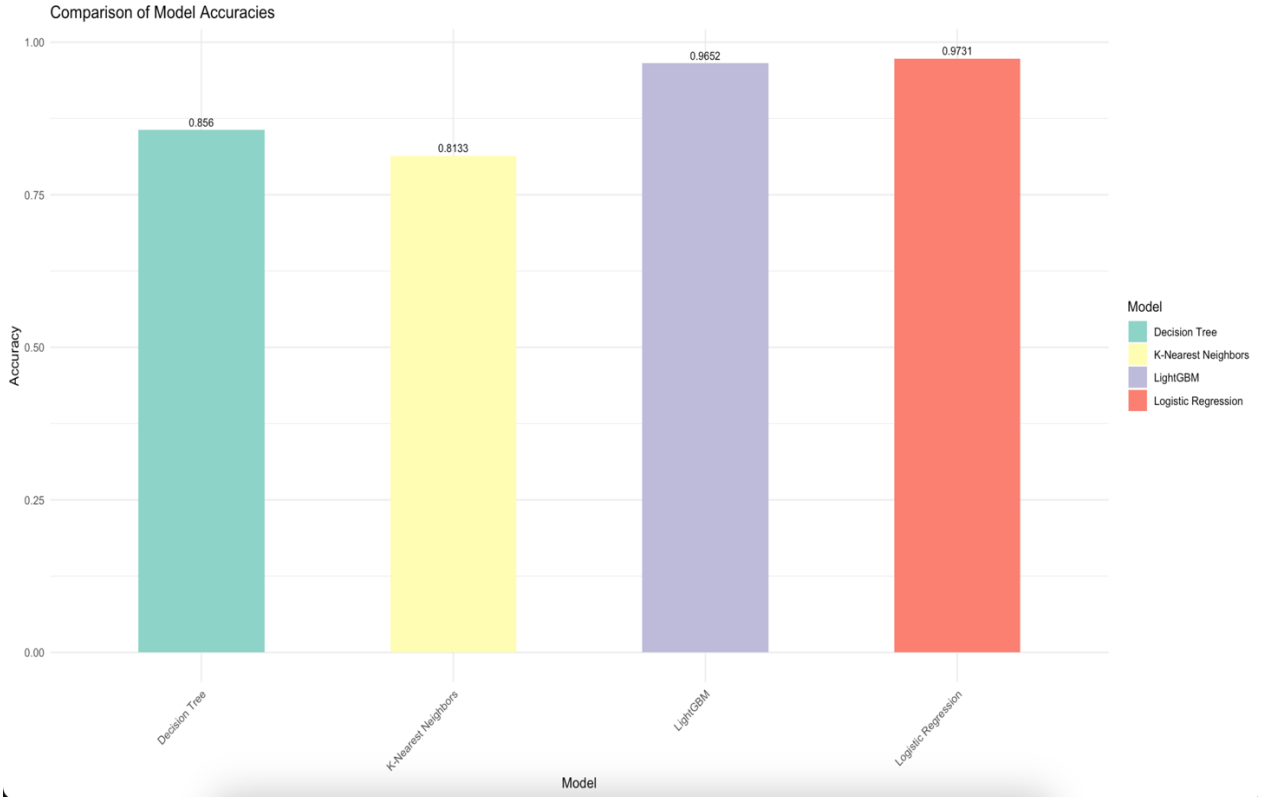
Accuracy: Logistic Regression has the highest accuracy (97.31%), followed closely by LightGBM (96.68%), then Decision Tree (85.60%), and KNN (81.33%). Accuracy measures the proportion of correctly predicted instances out of the total instances. A higher accuracy indicates better overall performance.

Precision: Logistic Regression also achieves the highest precision (97.22%), followed by LightGBM (96.53%), Decision Tree (86.24%), and KNN (80.26%). Precision measures the proportion of true positive predictions out of all positive predictions made by the model.

Recall: Logistic Regression and LightGBM have similar recall values, indicating their ability to identify most of the actual positive instances. Decision Tree follows, while KNN has the lowest recall among the methods.

F1 Score: Logistic Regression and LightGBM have comparable F1 scores, reflecting a balance between precision and recall. Decision Tree follows, and KNN has the lowest F1 score among the methods.

In summary, logistic regression performs the best overall in terms of accuracy, precision, recall, and F1 score for predicting obesity levels in this study.

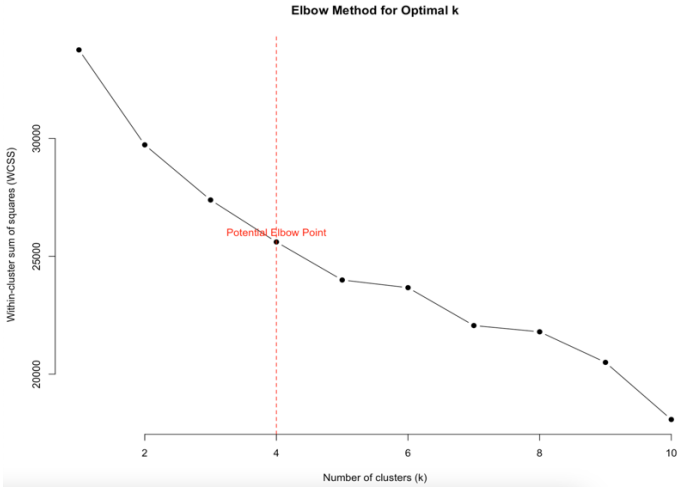


5.2. Unsupervised Learning Method

5.2.1. K-means

K-Means is a widely used unsupervised machine learning algorithm to separate data points into groups or clusters based on similarity. It divides the data into 'k' clusters, where each data point belongs to the cluster with the closest mean.

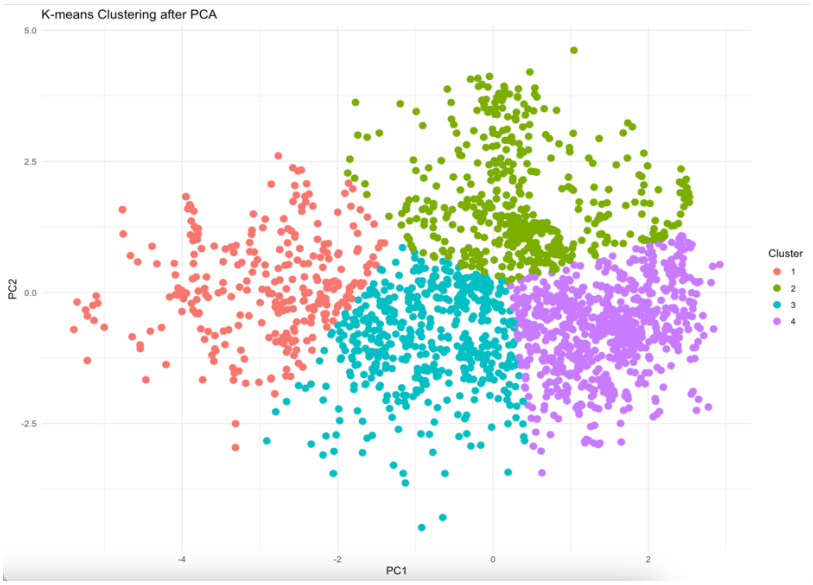
In this obesity study, K-Means identified 4 different groups of individuals based on their physical characteristics and eating habits.



Optimal k value was defined appliyng Elbow method. The graph illustrates that optimal k value is 4.

Table of k-means clustering results

Clusters	Target Variables							
.	1	2	3	4	5	6	7	
1	193	189	9	1	0	91	53	
2	0	21	110	97	1	58	94	
3	0	12	93	19	322	41	18	
4	79	65	139	180	1	100	125	



Cluster 1:

This cluster has 536 data points. Lower values for features such as Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE and CALC.Higher values for features like family_history_with_overweight, FAVC, and MTRANS.

Cluster 2:

This cluster has 381 data points. Moderate to high values for features like Age, Height, Weight, family_history_with_overweight, and FAF. Higher values for features like FAVC and FCVC.

Cluster 3:

This cluster has 505 data points. Lower values for features like Age, Height, and Weight. Higher values for features like family_history_with_overweight, CAEC, SCC ,and MTRANS.

Cluster 4:

This cluster has 689 data points. Moderate to high values for features like Age, Height, Weight, FCVC ,NCP , CH2O, and FAF. Higher values for features like family_history_with_overweight and MTRANS.

Conclusion

In conclusion, this study demonstrates the effect of machine learning techniques in predicting and understanding obesity levels. Supervised learning methods, particularly logistic regression, display remarkable performance in accurately classifying individuals into different obesity categories based on their attributes. Logistic Regression achieved an accuracy of 97.31%, precision of 97.22%, recall of 97.22%, and F1 score of 97.22%, indicating its superiority in obesity prediction. LightGBM and Decision Tree also performed well, though with slightly lower metrics. Additionally, the application of unsupervised learning via K-Means clustering identified four distinct segments based on similarities in their characteristics. These findings underline the potential of machine learning as a valuable tool for healthcare professionals in early detection and intervention of obesity. By leveraging machine learning, healthcare practitioners can enhance diagnostic and treatment outcomes, leading to more effective obesity management strategies and ultimately improving individual and societal health outcomes.