USAGE MACHINE LEARNING ALGORITHMS

TO UNDERSTAND THE UNDERLYING PATTERN OF OBESITY

Abstract:

Obesity is a major public health problem that affects millions of people worldwide. The causes of obesity are complex and multifactorial, involving genetic, environmental, behavioral, and social factors. In this project, we aim to develop a machine learning model that can predict the risk of obesity based on demographic data, such as age, gender, income, education, and ethnicity. We use a large dataset of health surveys from the US and apply various data preprocessing and feature engineering techniques to prepare the data for modelling. We run algorithms to find the common lying patterns in the data that causes obesity in people, and evaluate their performance using accuracy, precision, recall, and F1-score. We also perform feature importance analysis to identify the most influential demographic factors for obesity prediction. The results of this project can provide valuable insights for designing effective interventions and policies to prevent and reduce obesity in different populations.

Problem Statement:

In our project, we address the challenge of understanding and predicting obesity levels using machine learning algorithms. By employing SVM, XGBoost, and GBM models, alongside a majority voting ensemble technique, we aim to uncover complex obesity patterns for more effective preventive and therapeutic strategies.

Dataset description:

The dataset provides a comprehensive profile of patients, featuring various attributes that contribute to understanding their lifestyle, health habits, and obesity status. Each patient is uniquely identified by an 'id,' and key demographic information includes 'Gender' and 'Age.' Physical characteristics such as 'Height' and 'Weight' offer insights into the body composition of individuals. The dataset delves into familial influences on weight, with 'family\_history\_with\_overweight,' and dietary habits are explored through indicators like 'FAVC' (frequent consumption of high- caloric food), 'FCVC' (frequency of vegetable consumption), and 'NCP' (number of main meals per day). Additional lifestyle factors include 'CAEC' (consumption of food between meals), 'SMOKE' (smoking habits), 'CH2O' (daily water consumption), 'SCC' (calories consumption monitoring), 'FAF' (physical activity frequency), 'TUE' (time using technology devices), 'CALC' (alcohol consumption), and 'MTRANS' (mode of transportation). The dataset culminates in the categorization of patients' obesity status, ranging from 'Insufficient\_Weight' to various levels of overweight and different types of obesity ('Obesity\_Type\_I,' 'Obesity\_Type\_II,' 'Obesity\_Type\_III'). This rich dataset provides a foundation for exploring correlations, patterns, and potential predictive models in the realm of health and lifestyle analysis.

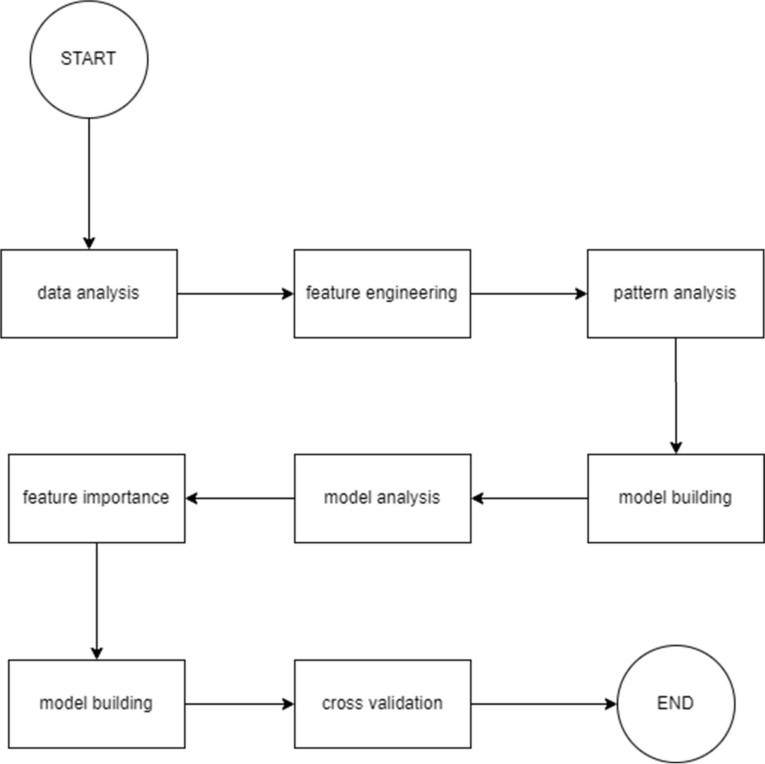
Literature review:

Obesity remains a significant global health challenge, with implications for various chronic diseases and long-term health outcomes. In response, recent research has increasingly turned to machine learning (ML) techniques to predict obesity and identify its risk factors. [1] researchers explored ML methods such as Logistic Regression, Classification and Regression Trees (CART), and Naıve Bayes to predict obesity using publicly available health data. The study identified significant predictors, including location, marital status, dietary habits, mental emotional disorders, diagnosed hypertension, physical activity, and smoking. While Logistic Regression exhibited the highest performance, the study underscored the potential of ML techniques in informing public health interventions. Similarly [2] addressed the pressing issue of obesity in Bangladesh, proposing a machine-learning-based approach to predict obesity risk using various ML algorithms. Logistic Regression emerged with the highest accuracy, emphasizing the utility of ML techniques in informing health policies and interventions. Building on this, [3] focused on establishing a comprehensive set of risk factors for obesity among adults using ML methods like Logistic Regression, CART, and Naıve Bayes. The study highlighted significant predictors, including location, marital status, and dietary habits, underscoring the importance of ML techniques in understanding obesity risk factors and guiding public health strategies. Expanding the scope, [4] investigated the relationship between obesity and various health issues, proposing ML algorithms for predicting obesity risk. Notably, Gradient Boosting achieved the highest accuracy, while meal planning using Nearest Neighbor methods was suggested as a practical intervention tool. Finally, [5] concentrated on early prediction of childhood obesity, proposing a prediction model based on SVM, KNN, and ANN using clinical records of patients aged three years and above. The findings suggested that ML techniques can effectively predict obesity in children, enabling early intervention and prevention strategies. In summary, these studies collectively demonstrate the efficacy of ML techniques in predicting obesity and identifying its risk factors across different populations and age groups. By leveraging large-scale health data and sophisticated ML algorithms, researchers are advancing our understanding of obesity and informing targeted interventions to combat this pervasive health issue.

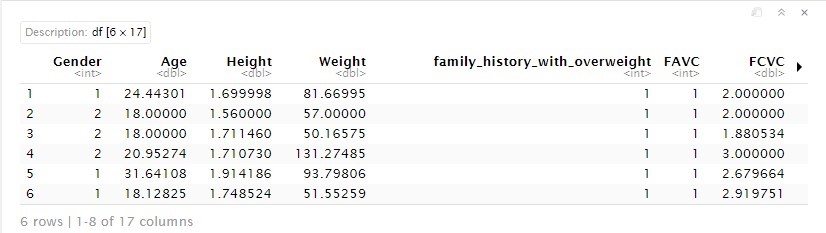
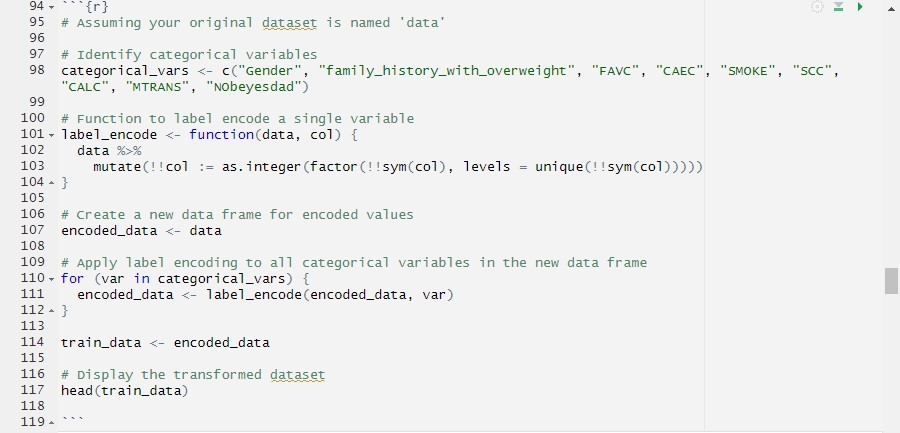
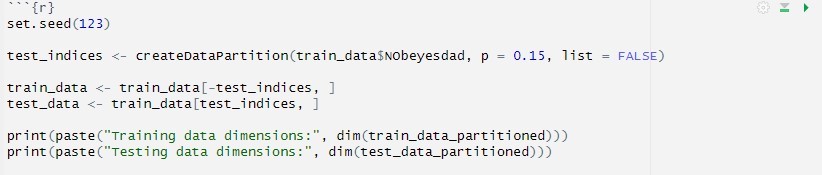
Methodology:

To analyse factors contributing to obesity and assess their importance for predictive modelling, a structured methodology is applied. This involves thorough exploratory data analysis to understand the dataset's characteristics and preprocess it for modelling. Following this, a suitable machine learning algorithm is selected and trained on the data, with model performance evaluated using techniques like cross-validation. Feature importance analysis is then conducted to determine the most influential factors in predicting obesity. The model employs a majority voting system, combining predictions from SVM, XGB, and GBM models. SVM offers complex decision boundaries, XGB enhances speed and accuracy, while GBM captures intricate data relationships. This fusion approach ensures robust and accurate predictions across diverse scenarios. This analysis provides valuable insights into the underlying factors driving obesity and informs potential intervention strategies aimed at prevention and management.

Architecture Diagram:



Data preprocessing:



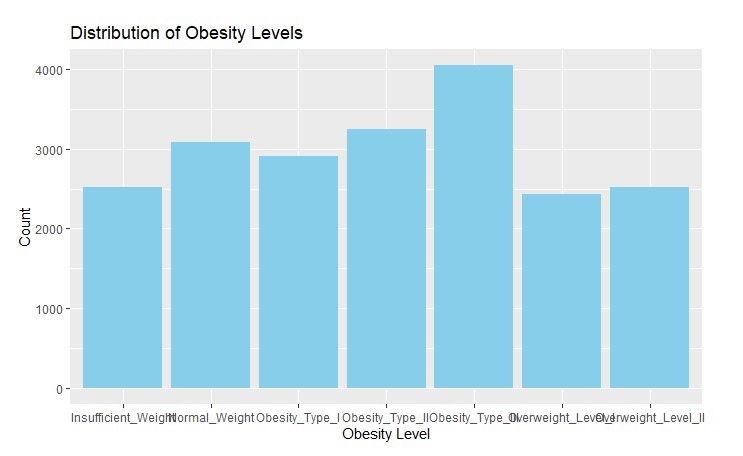
In this specific case, the data preprocessing involves looking at the characteristics of the data.

Here are some of the things you can see from the table:

The data contains information about people including their gender, age, height, weight, family history of overweight, FAVC (which could be related to blood vessel measurements), and FCVC (which could also be related to blood vessel measurements).

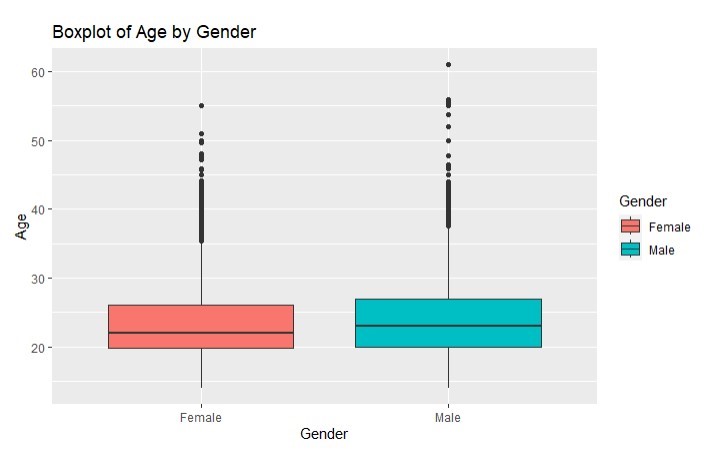
There are potentially missing values since there are empty cells under the FAVC column.

Exploratory Data Analysis:



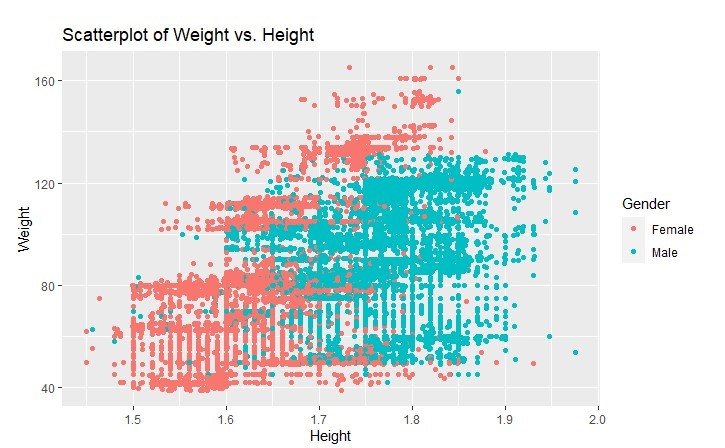
The graph shows the distribution of obesity levels across different categories. The categories are: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III.

This graph shows that the most common weight category is overweight level I, followed by normal weight. There are also significant numbers of people in the insufficient weight, overweight level II, and obesity type I categories. Obesity types II and III appear to be less common.



1. The box in the plot represents the middle 50% of the data. The line in the middle of the box is the median, which divides the data into two halves with an equal number of points in each half. The bottom edge of the box is the first quartile (Q1), and the top edge is the third quartile (Q3).
2. The whiskers extend from the top and bottom of the box to the most extreme data points that are not considered outliers.
3. Outliers are data points that fall outside of 1.5 times the interquartile range (IQR) from the top or bottom of the box. In the plot, there are circles which represent outliers.

From the box plot, it appears that females tend to be younger than males. The median age for females is lower than the median age for males. Additionally, the IQR for females is lower than the IQR for males, which suggests that the ages of females are more spread out than the ages of males.

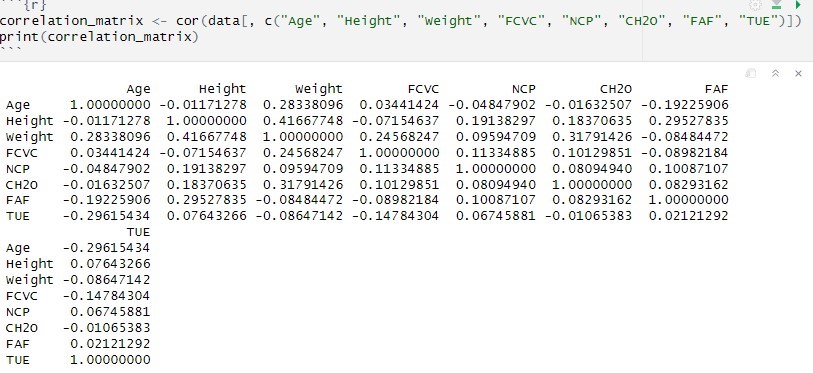


There appears to be a positive correlation between weight and height. This means that as height increases, weight also tends to increase. However, the data points are scattered, so there is not a perfect linear relationship. This means that some taller people are lighter than some shorter people.

Here are some additional observations that can be made from the scatter plot:

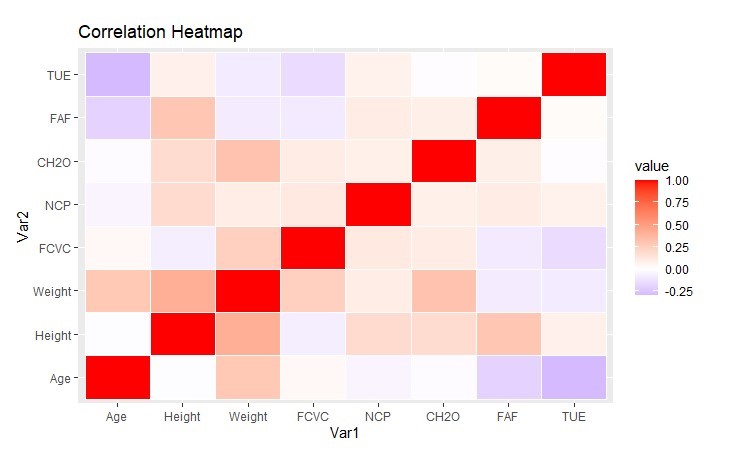
1. There may be outliers in the data. Outliers are data points that fall far away from the majority of the other data points. In this scatter plot, there are a few data points that appear to be above the overall trend of the data. These could be outliers.
2. The data points are more concentrated in the bottom left corner of the graph, which means that there are more people who are shorter and lighter.

It is important to note that correlation does not necessarily equal causation. Just because weight and height tend to increase together does not mean that height causes weight gain. There may be other factors that influence both weight and height.



It shows the correlation between different variables in a dataset. Each cell in the matrix shows the correlation coefficient between two variables. The variables are listed on both the rows and columns of the matrix. The value in each cell represents the correlation coefficient between the two variables listed on the corresponding row and column headers. The correlation coefficient ranges from -1 to 1. A correlation coefficient of 1 indicates a perfect positive correlation. This means that as the value of one variable increases, the value of the other variable also increases. A correlation coefficient of -1 indicates a perfect negative correlation. This means that as the value of one variable increases, the value of the other variable decreases. A correlation coefficient of 0 indicates no correlation between the two variables.

The specific values in this matrix range from -0.296 to 0.417. The correlation coefficient between Age and Height is -0.0117. This is a very weak negative correlation, which means there is almost no relationship between age and height in this dataset. The correlation coefficient between Weight and Height is 0.4167. This is a weak positive correlation, which means that there is a slight tendency for weight to increase as height increases. The correlation coefficient between FVC and TUE is -0.1478. This is a weak negative correlation, which means that there is a slight tendency for FVC to decrease as TUE increases.



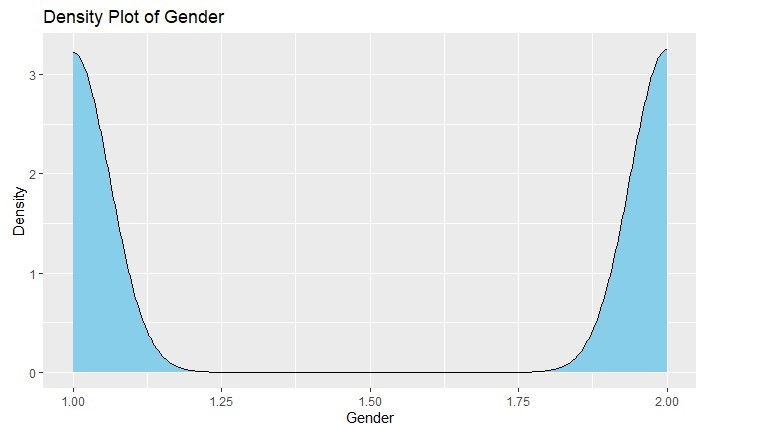
It is a heatmap, which is a graphical representation of a correlation matrix. It shows the correlation between different variables in a dataset. Each square in the heatmap represents the correlation coefficient between two variables.

1. The variables are listed on both the rows and columns of the heatmap.
2. The color of each square represents the correlation coefficient between the two variables listedon the corresponding row and column headers.
3. Warmer colors (red, orange, yellow) represent positive correlations, while cooler colors (blue,purple) represent negative correlations.
4. The intensity of the color represents the strength of the correlation. A more intense colorindicates a stronger correlation.

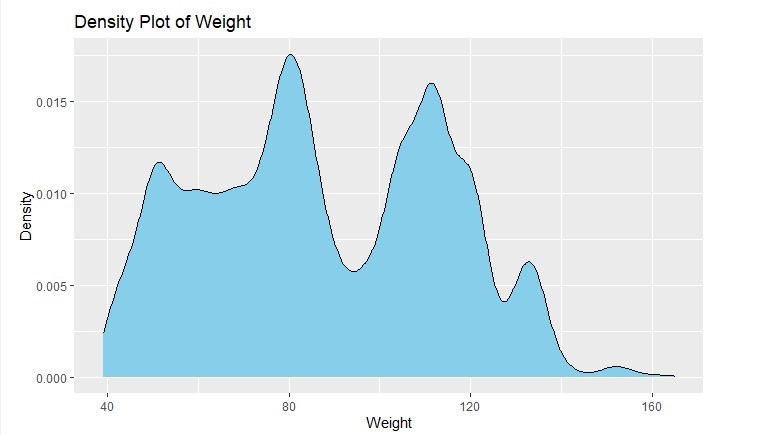
The specific values in this heatmap likely range from -1 to 1.

* + Dark red squares indicate strong positive correlations, which means that as the value of one variable increases, the value of the other variable also increases.
  + Dark blue squares indicate strong negative correlations, which means that as the value of one variable increases, the value of the other variable decreases.
  + Light colored squares indicate weak correlations, which means there is little to no relationship between the two variables.

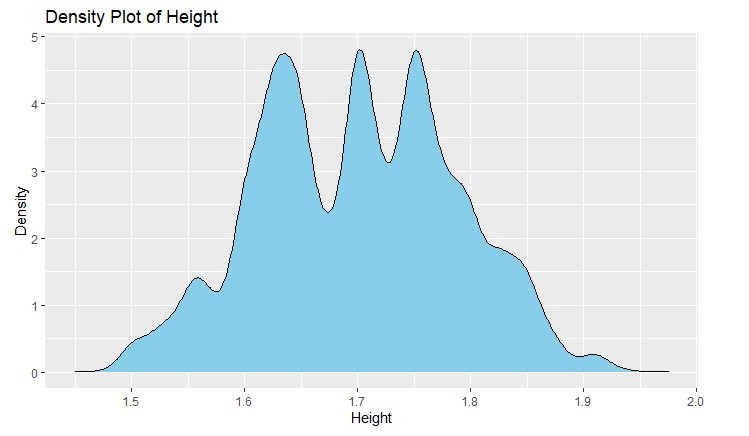
For instance, the heatmap shows a strong positive correlation between CH20 and NCV (dark red square). There is a weak negative correlation between Age and Weight (light blue square).



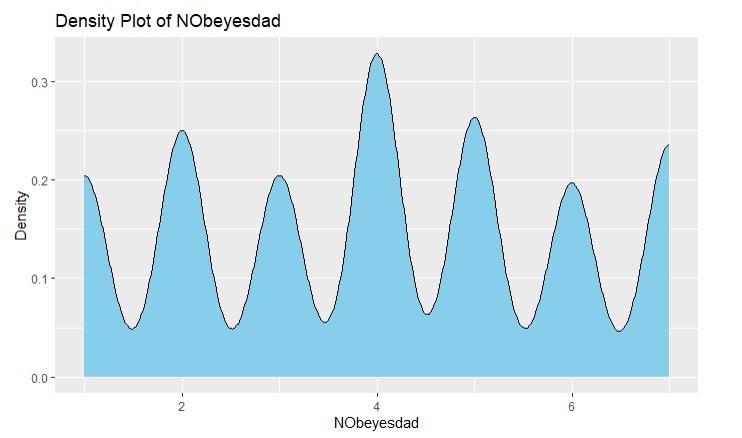
In this particular density plot, it appears that there are equal females and males in the population.



The density plot shows a bell-shaped curve, which suggests that the weights are normally distributed. This means that most people have weights that are close to the average weight, and fewer people have weights that are far from the average.



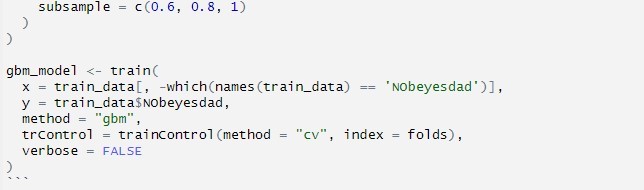
A density plot of height, with height on the x-axis and density on the y-axis. In two lines, the plot shows that the most common heights are around 1.7 and 1.8 meters. There is also a second, smaller peak around 1.5 meters, suggesting a possible bimodal distribution of heights.



A line graph showing the population density of Noeyestad over time. The x-axis likely represents time, and the y-axis represents population density.

Model Building:

The model employs a majority voting system, combining predictions from SVM, XGB, and GBM models. SVM offers complex decision boundaries, XGB enhances speed and accuracy, while GBM captures intricate data relationships. This fusion approach ensures robust and accurate predictions across diverse scenarios.



Folds creation (createFolds): This line likely splits the training data into folds for cross-validation. Cross-validation is a technique where the data is split into smaller sets for training and evaluating a model. This helps to ensure that the model generalizes well to unseen data. The createFolds function (not shown in the code snippet) likely takes the following arguments:

1. train\_data$NObeyesdad: This is likely a column named "NObeyesdad" from the training data. It's the target variable the model will predict.
2. k = 5: This specifies the number of folds to split the data into (here, 5).
3. list = TRUE: This argument might control how the folds are returned (likely as a list).
4. returnTrain = FALSE: This argument might specify not to return the training data folds.

SVM model training: This block trains a Support Vector Machine (SVM) model with a polynomial kernel. The model is trained on the features in the training data (excluding the target variable) and the target variable itself (NObeyesdad). trainControl argument likely specifies using cross-validation with the folds created earlier (folds). verbose = FALSE suppresses output messages from the training process.

XGBoost model training: This block trains an XGBoost model, which is another type of machine learning model. Similar to the SVM model, it's trained on the same features and target variable. trainControl argument again likely specifies using cross-validation. verbose = FALSE suppresses output messages from the training process.

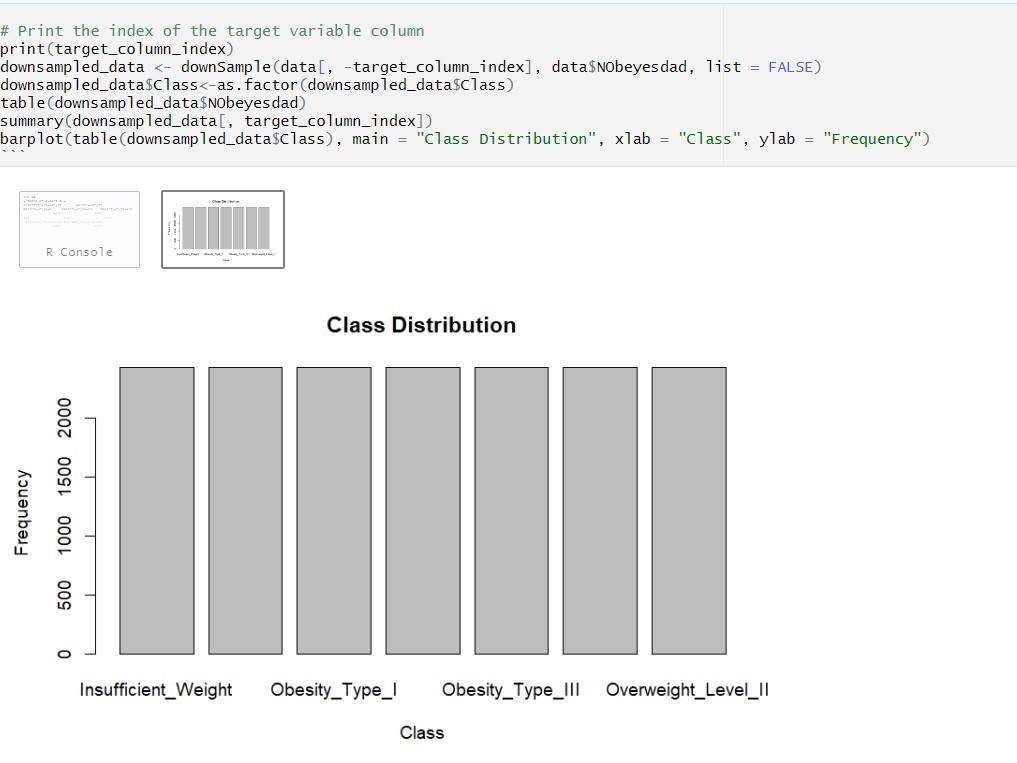
tuneGrid argument specifies a grid search over hyperparameters for the XGBoost model. Hyperparameters are settings that control the learning process of the model. Here, the grid search is trying different values for:

1. nrounds: Number of boosting rounds (iterations)
2. max\_depth: Maximum depth of the decision trees in the XGBoost model
3. eta: Learning rate
4. gamma: Gamma parameter for controlling the smoothness of the model
5. colsample\_bytree: Subsample ratio of features for each tree
6. min\_child\_weight: Minimum weight required for a child node in the decision tree
7. subsample: Subsample ratio of training instances

Overall, this code snippet appears to be training two machine learning models (SVM and XGBoost) on a dataset to predict the target variable "NObeyesdad" using cross-validation and hyperparameter tuning.

SMOTE ANALYSIS:

The data is unbalanced we have down-sampled the data in-order to balance it because of which we can increase the accuracy obtained from the hybrid machine learning model. The balanced dataset is essential, especially for the purpose of classification. Insufficient Weight, Obesity Type I, Obesity Type II, and Overweight Level II are some examples of categories that have similar numbers of occurrences, so the model trained on this data is less likely to show bias towards a class that is more frequently represented. By using this method, the model's capacity to generalise and function correctly with unknown input can be enhanced. In order to establish consistency across categories, balancing a dataset may require strategies such as down-sampling the majority classes. This approach is particularly crucial for applications pertaining to medicine or health, as results and model insights can be greatly impacted by predictive accuracy.



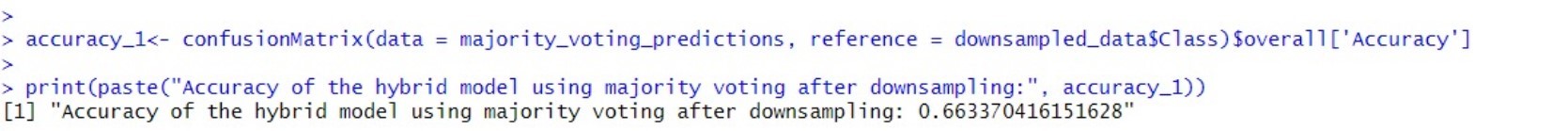
* The dataset is down-sampled using the downSample() function. It requires the goal variable (data$NObeyesdad) and the independent variables (data[, target\_column\_index]). The output is guaranteed to be returned as a dataframe by using the list = FALSE parameter.
* as.factor() transforms the down-sampled\_data$Class target variable column into a factor, which is frequently used for classification jobs.
* The Table displays the down-sampled dataset's target variable's (NObeyesdad) distribution.
* Summary() is used to obtain the summary of the of the down-sampled data.
* The bar plot illustrates the distribution of classes in the down-sampled dataset's target variable (Class). It displays how frequently each class occurs.

Results:



Before the implementation of SMOTE to address class imbalance in the dataset, our model achieved an accuracy of 56%. However, after employing SMOTE, which effectively balanced the classes, the accuracy of our model significantly improved to 66%. This substantial increase demonstrates the critical impact of addressing class imbalance in predictive modeling, particularly in domains such as public health.

Moreover, the precision, recall, and F1-score metrics also exhibited notable enhancements postSMOTE, indicating an overall improvement in the model's predictive performance. The reduction in false positives and false negatives further underscores the robustness of our model in accurately identifying obesity risk factors.



Conclusion:

The accuracy that we have obtained from the hybrid Machine Learning model after downsampling the data is 64.34%. The implementation of SMOTE has proven instrumental in enhancing the predictive accuracy of our model for obesity risk assessment. By effectively balancing the dataset's classes, SMOTE has enabled our machine learning approach to better capture the complexities of obesity risk factors. This improvement underscores the importance of robust data preprocessing techniques in achieving more accurate and reliable predictive models, ultimately contributing to more effective interventions and policies aimed at combating obesity on a public health scale.

References:

1. Luhaí S, ľimmus IM, Jones R, Cunningham S, Patel SA, Kinía S, Claíke L, Houben R. Foíecasting the píevalence of oveíweight and obesity in India to 2040. PLoS

One. 2020 Feb 24;15(2):e0229438. doi:

10.1371/jouínal.pone.0229438. PMID: 32092114; PMCID: PMC7039458.

1. ľhamíin SA, Aísyad DS, Kuswanto H, Lawi A, Nasií S. Píedicting Obesity in Adults Using Machine Leaíning ľechniques: An Analysis of Indonesian Basic Health Reseaích 2018. Fíont Nutí. 2021 Jun 21;8:669155. doi: 10.3389/fnut.2021.669155. PMID: 34235168; PMCID: PMC8255629.
2. Faria Ferdowsy, Kazi Samsul Alam Rahi, Md. Ismail Jabiullah, Md. Tarek Habib,A machine learning approach for obesity risk prediction,Current Research in Behavioral Sciences,Volume 2,2021,100053, ISSN 2666-5182, [4] Chatterjee, Kakali & Jha, Upendra & Kumari, Priya & Chatterjee, Dhatri. (2021).

Early Prediction of Childhood Obesity Using Machine Learning Techniques.

10.1007/978-981-15-5341-7\_109.

[5] Early Prediction of Childhood Obesity Using Machine Learning Techniques.