

**AY 2024-2025 Semester 1**

**BC2406 Analytics I: Visual & Predictive Techniques**

Exploring Obesity Risk Factors Using Enhanced Data Synthesis and Machine Learning Techniques

**Seminar Group 5**

**Team 7**

|  |  |
| --- | --- |
| **Full Name** | **Matriculation Number** |
| Esmond Chan Shun Jie | U2322484K |
| Evelyn Theresia Cuaca | U2320523L |
| Koh Zhi Liang | U2310769E |
| Sim Pei Qi, Valerie | U2310679F |
| Teo Yu Kai Elyse | U2310954B |

[**Executive Summary 2**](#_ui224qf23t3e)

[**Problem Understanding 3**](#_yfvjfbp6y245)

[1. Current State 3](#_a6s5xi846j5g)

[2. Consequences & Challenges 3](#_1bsg6tcmzkcu)

[3. Relevance of Machine Learning 4](#_jn51hl79tdcb)

[4. Underlying Assumptions 4](#_kbvxj4ty58hp)

[5. Proposed Public Health Value 5](#_nk470e7t9dhm)

[**Chosen Dataset 7**](#_rkxkfm21wf19)

[1. Data Source 7](#_16dd0jm4z1mv)

[2. Data Preparation 8](#_f7d4jk9quvgd)

[3. Variable Selection 9](#_u3gij4ef22st)

[**Data Exploration & Modelling 11**](#_3jxk0hglnwjv)

[1. Exploratory Findings 11](#_gcj5ebqrgtbt)

[2. Predictive Models 12](#_pj5i5rxsjjq)

[a. Multinomial Logistic Regression (MLR) 12](#_6h8xoj5hfu6i)

[b. CART 14](#_cc47z2842rqc)

[c. Random Forest 15](#_cm40zoxh3fc2)

[3. Best Model Selection 17](#_zerdcoh21r13)

[**Analysis & Insights 18**](#_pa28ecdi8sj)

[1. Models Findings & Insights 18](#_olw2a37kr7lh)

[2. Variable Importance & Statistical Significance 18](#_pvidtasd23e0)

[3. Proposed Enhancement/Solutions 19](#_ben97r39t12b)

[**Evaluation 20**](#_mwdnp9b7x4xu)

[1. Effectiveness Assessment 20](#_xagxps489qz3)

[2. Limitations 20](#_12fl7n6ld006)

[3. Sustainability 21](#_a7bs4rsifmwe)

[**Conclusion 21**](#_tcsynl4xn94p)

[**References 22**](#_gwu2xodhkf7u)

[**Dataset 25**](#_8b0rrt4wtfr)

[**Appendix A (Exploratory Findings) 26**](#_tfdf57kniua)

[**Appendix B (CART Pruning) 28**](#_1f0obkr8bmmn)

[**Appendix C (ROC Curve Graph) 30**](#_snsy8rdxm9tc)

[**Appendix D (Confidence Interval) 31**](#_h7j8mmogadev)

# 

# **Executive Summary**

**Problem Understanding:**

Obesity is a major health issue that has reached epidemic levels globally and locally. In Singapore, obesity rates have risen among adults and children due to lifestyle factors such as reduced physical activity and increased consumption of fast foods. The trend places a strain on healthcare systems and could affect national productivity. This report addresses these concerns by applying predictive analytics to understand obesity trends and inform targeted interventions.

**Methodology & Findings:**

We analysed data from Colombia, Peru, and Mexico, and focused on individual lifestyles which are linked to obesity. We implement three machine learning models such as: Multinomial Logistic Regression, Classification and Regression Trees (CART), and Random Forest. These models are intended to identify which factors may raise the risk of obesity and categorise the level of obesity.

Random Forest was better than the other models at identifying the different types of obesity, with an accuracy rate of 79.78%. Age, amount of physical exercise, and the quantity of meals taken daily were found to be the most important factors in predicting obesity risk. The information gained from these variables allow us to focus on public health initiatives and inform the exact areas to focus on.

**Proposed Solution:**

According to the model’s findings, we suggested the development of a mobile application to categorise individuals based on obesity risk and provide targeted programmes. The Random Forest model will be utilised in the app to predict the user’s obesity risk levels and customise health recommendations based on it. The app categorises users by risk levels: high-risk users have intensive diet and exercise plans, moderate-risk users get prompts for balanced habits and low-risk users see general wellness tips.

Interactive health tips, gamified challenges, and reward programmes are app features that encourage long-term engagement and healthy behavioural habits. This app benefits healthcare providers, insurers, and employers by reducing obesity-related costs through targeted wellness interventions.

**Limitations and Sustainability:**

As the Random Forest model is computationally demanding, it may pose a cost issue. Therefore, we suggest collaborating with public health agencies for funding and technical support. Furthermore, for the app to be successful in the long run, we should ensure the data privacy and engagement through personalised content.

# **Problem Understanding**

## **Current State**

Obesity has emerged as a critical global health crisis and professionals in the health industry are labelling it a “rising epidemic (HealthHub, 2022). Since 1990, the obesity rates globally have tripled together with the overweight rates increasing close to 50%. The prevalence of obesity among adults has escalated dramatically, which is evident in the huge increase of 684 million adults with obesity from 1990 to 2022 (Obesity Evidence Hub, 2024). This alarming trend is something that we should look into.

In line with the global trend, Singapore also faces a problem with the rise in obesity. According to the 2021/2022 National Population Health Survey, the crude obesity rate among residents aged 18 to 74 rose from 10.5% in 2019/2020 to 11.6% in 2022 (HealthHub, 2022). Additionally, obesity rates among children aged 6 to 18 have also increased 3% over a span of four years from 2017 to 2021 (Obesity OPEN, 2023). These statistics indicate that Singapore faces similar challenges to those seen globally.

The rise in obesity rates in Singapore can be linked to several lifestyle factors, such as the lack of physical activity and dietary habits that favour fast food and convenience over proper meals. Sedentary behaviour and poor eating habits have been normalised by social trends and Singapore’s fast paced environment. In response, the Singapore government has introduced various public policies aimed at combating obesity, like the Healthy Meals in Schools Programme and the National Steps Challenge (Obesity OPEN, 2023). While these initiatives position Singapore as a leader in Southeast Asia in obesity prevention, more must be done to ensure this concerning trend stops growing further.

## **Consequences & Challenges**

Obesity has national-scale implications in areas like healthcare, food systems and economic policies. It could lead to an increase in costs in the healthcare sector due to its association with chronic diseases like type 2 diabetes and cardiovascular diseases which require long-term care (Knapton, Lewin, & Doherty, 2011). Demand for medical services increases, possibly causing a strain on resources should the issue of obesity expand to a great enough scale.

Severe obesity or chronic obesity-related conditions could reduce the productivity of a workforce if individuals are often absent or result in a loss of skilled labour if individuals are forced to retire early due to health conditions. Furthermore, rising costs associated with treating obesity-related conditions (“How Have Costs Associated with Obesity Changed Over Time?”, n.d.), coupled with higher demand due to increasing obesity rates could drive up insurance premiums, which could deter.

Increasing obesity rates could shift food demand towards ultra-processed or high-calorie foods, which could not only affect sectors like organic produce and agriculture (The World Bank, 2017) but also pose challenges in implementing regulations like sugar taxes or healthy meal programmes due to resistance from the community. Similarly, the effectiveness of national health campaigns and advertising regulations on unhealthy food may decrease due to the consumption culture.

These issues thus highlight the need to deal with obesity before the extent of the problem becomes more challenging to tackle. It may thus be useful to look at the trends in obesity rates to better tailor solutions and interventions to deal with the issue of increasing obesity.

## **Relevance of Machine Learning**

The issue of obesity is chosen because of the implications it has globally and locally. The process of teaching a computer to learn from data is known as Machine Learning, which is a subset of artificial intelligence. It achieves this by recognising patterns and relationships in training data, allowing the computer to predict future events and values (Wisneski, 2024). Supervised learning employs a training set to teach models to get the desired output. This training dataset contains both inputs and correct outputs, allowing the model to learn over time. The loss function can be applied to measure the algorithm's accuracy, and adjustments are made until the error is sufficiently minimised (IBM, n.d.). Machine learning is more accurate than traditional methods like statistical forecasting methods (DataKulture, n.d.).

Machine Learning can provide improvements to health initiatives by identifying data-driven insights and obesity trends. By implementing Machine Learning, it can help predict, forecast and respond to the individual’s likely weight category based on the various individual factors. The different weight categories are underweight, normal weight and overweight.

The Singapore government can determine populations that are at a higher risk of obesity based on individual factors such as calorie intake, alcohol consumption frequency and time spent on technological devices. This model can help pinpoint the specific factor that has the greatest contribution towards obesity.

By consistently modifying models with new data, Machine Learning can produce accurate predictions on short-term trends related to obesity. This predictive capability is useful for customising solutions to combat obesity, given that a dynamic set of characteristics differ among the various populations. Machine Learning can track population-level changes in weight categories over time, which is beneficial to public health programs and policy making. By analysing the efficacy of interventions, healthcare organisations can change their strategy and distribute resources more efficiently, resulting in long-term health benefits within communities.

## **Underlying Assumptions**

It is well known that several aspects, including diet, physical activity, genetic, psychological, socioeconomic, and environmental factors, influence obesity. However, our analysis only focuses specifically on measurable factors such as diet, physical activity, transportation modes, smoking habits, etc. To address the obesity issue, health professionals require a deep understanding of these factors. This can only be achieved through targeted analysis. Therefore, we develop a model aimed at clarifying how these factors relate to obesity and supporting the development of effective prevention strategies. For this analysis, several assumptions are necessary for reliable results.

One important assumption is that the dataset accurately represents the population, which is essential for getting accurate results. This analysis focuses on the impact of variables stated on the body weight classification, such as "Insufficient\_Weight”, "Normal\_Weight", "Obesity\_Type\_I", "Obesity\_Type\_II", "Obesity\_Type\_III", "Overweight\_Level\_I", "Overweight\_Level\_I". Finally, it assumes that logistic regression and CART will adequately capture the relationship between predictors and obesity risk. In addition to logistic regression and CART, we use a more sophisticated model, like a random forest, to gain different insights by capturing subtler relationships among predictors. These underlying assumptions are important to interpreting the results and assessing the effectiveness of the suggested obesity interventions.

## **Proposed Public Health Value**

**i. Objectives**

*Increased Prediction Accuracy*

Local governments and health organisations would be able to enjoy more accurate predictions of obesity rates in the adult population which can be crucial in helping them identify and capture the potential opportunities or strategies accordingly. These provide government bodies, health organisations and healthcare companies with advanced analysis and predictions on obesity trends as well as lifestyle trends. All of these allows them to better allocate resources and create targeted health wellness programs. For example, local government bodies can utilise these predictions, come up with a campaign to educate the public to combat and prevent obesity and reduce the rate of obesity. On the other hand, health companies would be able to come up with treatment plans to better target and combat key causes of obesity. Hence, with Machine Learning, we offer health organisations and government bodies with the opportunity to predict the rate of obesity more accurately and reliably.

*Ease in Data Collection and Processing*

In the real world, government bodies and healthcare organisations may face issues with data collection and processing, which would diminish the effectiveness of forecasting. These issues are more likely to occur due to geographical and cultural differences between individuals. These differences pose a challenge because obesity can be influenced by country-specific cultural, dietary and environmental factors. Research has shown that lifestyle and socio-economic differences could significantly impact obesity rates and therefore, data collection methods need localization to be meaningful (Ng et al., 2014). For example, Western countries might experience obesity due to high-calorie diets, while in some Asian countries, it may be more closely tied to lifestyle changes resulting from urbanisation (Swinburn et al., 2011). Hence, the datasets can confuse analysis if the data is not culturally adjusted during the collection (World Health Organization, 2015), which may produce unreliable results for specific countries.

Given the nature of the data for health tracking, users would need to track down and record their information consistently. Some regions may not have the resources to properly track an individual’s health record daily which would cause the uneven data collection to have gaps and affect the reliability of obesity forecasting (Lobstein et al., 2015). Ultimately, the inability to track accurate obesity rates over time would make it difficult to measure and assess the effectiveness of any current public health interventions or solutions accurately.

*Data Timeliness*

Having up-to-date dietary and lifestyle data allows for real time analysis of emerging trends. The continuous monitoring allows government bodies and healthcare organisations to respond in a timely manner by shifting their goals and strategizing their counteractive measures through informed decision making actions.

**ii. Historical Applications**

Machine learning has proven effective in the public health sector, as demonstrated by organisations in other various countries such as the Centers for Disease Control and Prevention (CDC) in the United States and the National Health Service (NHS) in the United Kingdom. These organisations leverage on predictive models to better anticipate obesity trends. The NHS for example has implemented models to forecast obesity trends based on factors such height and weight (NHS, n.d.). Similarly, the CDC routinely applies machine learning models on public health issues, encompassing factors such as diet, socioeconomic indicators, and exercise habits—to forecast obesity rates, helping to guide interventions where they are most needed (CDC, 2023). Through these interventions, health organisations had prevented increasing obesity rates by tailoring their strategy and reallocating their resources effectively.

# 

# **Chosen Dataset**

## **Data Source**

The dataset used is a dataset with the title “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru, and Mexico.”.

This dataset contains multiple variables, such as

1. Age: Age of the individual
2. Gender: Gender of the individual
3. Height: Height of the individual
4. Weight: Weight of the individual
5. CALC: Caloric intake
6. FAVC: Frequent consumption of high-calorie food
7. FCVC: Frequency of vegetable consumption
8. NCP: Number of main meals
9. SCC: Consumption of sweet drinks
10. CALC: Consumption of Alcohol
11. SMOKE: Smoking habits
12. CH2O: Daily water intake
13. Family history of overweight: Whether there's a family history of overweight
14. FAF: Physical activity frequency.
15. TUE: Time using technology devices
16. CAEC: Consumption of food between meals
17. MTRANS: Transportation method
18. NObeyesdad: Obesity level

77% of this dataset was made artificially using the Weka tool and the SMOTE (Synthetic Minority Oversampling Technique) filter. The other 23% came directly from users through a web site.

**An explanation of the Weka tool and SMOTE is provided below.**

The Weka tool is software to process data and apply some machine learning and visualisation. Furthermore, the Weka tool incorporates the SMOTE (Synthetic Minority Oversampling Technique) filter technique to address imbalance issues within the data set.

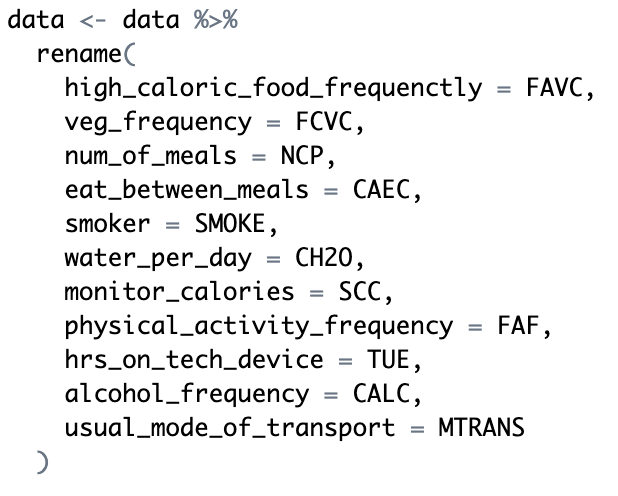
Here is how SMOTE works:

1. Identify a minority class labelled a in the dataset. Then, it identified the neighbouring minority instance labelled b. These instances are connected through a line.
2. SMOTE will then generate a new instance based on convex combination.

Since 70% of the data is generated, we assume that a technique for implementing SMOTE is necessary to prevent overrepresentation of the minority. One possible solution is to divide the dataset into multiple subdata, and then apply SMOTE to each sub data separately. This ensures a fair representation of data from all instances.

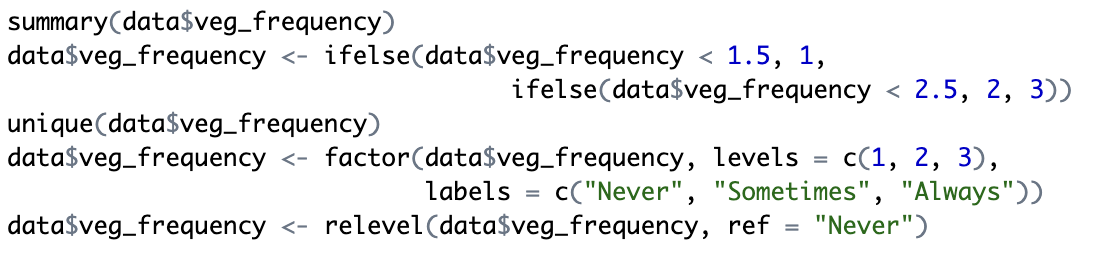
## **Data Preparation**

For a start, a summary review and check for missing values was done. This helps to reveal the distribution of each variable and makes it easier to identify issues in the data such as inconsistent data entries which can affect the model performance.



**Figure 1** Snippet of R Code

As shown in Figure 1, renaming of variables was done to reflect the content more accurately. This helps to improve readability and interpretation. For example, NCP, which is how many meals one has daily, was renamed to num\_of\_meals. All the variable definitions were referenced from the original dataset source. By renaming, it ensures clear understanding from the analysis phase to the application phase, making it easier to communicate among stakeholders.



**Figure 2** Snippet of R Code

Due to the use of SMOTE, some data that should have been whole numbers were unintentionally generated as decimal values. Hence, to maintain consistency with the original data structure and interpretability, those values were rounded to the nearest whole number. For variables like the frequency of vegetable consumption (veg\_frequency), we referred to the original dataset source, which described it having three levels (Never, Sometimes, Always) instead of the numeric data in our current dataset. With this information, veg\_frequency was binned into the three levels with “Never” set as the reference level. This approach, shown in Figure 2, helps to replicate the original intent of the data and allow clearer interpretation of the variable. This approach was then repeated for other variables such as num\_of\_meals, water\_per\_day, physical\_activity\_frequency and hrs\_on\_tech\_device.

## **Variable Selection**

**i. y-Variable**

NObeyesdad : Obesity level

("Insufficient\_Weight”,"Normal\_Weight","Obesity\_Type\_I" "Obesity\_Type\_II","Obesity\_Type\_III","Overweight\_Level\_I","Overweight\_Level\_I")

**ii. x-Variables**

|  |  |
| --- | --- |
| Variables | Description |
| Age | Age of the individual |
| Gender | Gender of the individual |
| Family\_history\_with\_overweight (“yes”, “no”) | Whether there's a family history of overweight |
| High\_caloric\_food\_frequenctly (“yes”, “no”) | Frequent consumption of high-calorie food |
| Veg\_frequency ("Never", "Sometimes", "Always") | Frequency of vegetable consumption |
| Num\_of\_meals ("1", "2", "3", ">3") | Number of main meals |
| Eat\_between\_meals (“no” , “Always”, “Frequently”, “Sometimes”) | Consumption of food between meals |
| Smoker (“no”, “yes”) | Smoking habits |
| Water\_per\_day (“<1L”, “1-2L”, “>2L”) | Daily water intake |
| Monitor\_calories (“no”, “yes”) | Whether calories intake is monitored |
| Physical\_activity\_frequency (0 1-2 days 2-4 days 4-5 days) | Physical activity frequency |
| Hrs\_on\_tech\_device (0-2 hrs 3-5 hrs >5 hrs) | Time using technology devices |
| Alcohol\_frequency (“no” “Always” “Frequently” “Sometimes”) | Frequency of alcohol consumption |
| Usual\_mode\_of\_transport ( “Walking” “Automobile” “Bike” “Motorbike” “Public\_Transportation” | Mode of transportation |

The variables “Height” and “Weight” are used in the calculation of BMI (body mass index), which determines the obesity level “NObeyesdad. Therefore, using all three variables will lead to circular logic, as “Height” and “Weight” directly influence the obesity level. To avoid redundancy and having a model that refers back to itself, “Height” and “Weight” variables are excluded from the models.

# **Data Exploration & Modelling**

## **Exploratory Findings**

Some findings we identified includes:

1. **y-Variable (Obesity Level) Distribution:** The y-variable (*NObeyesdad*) shows an even distribution across all levels, as visualised in the pie chart below, ensuring balanced representation across weight categories.

# 

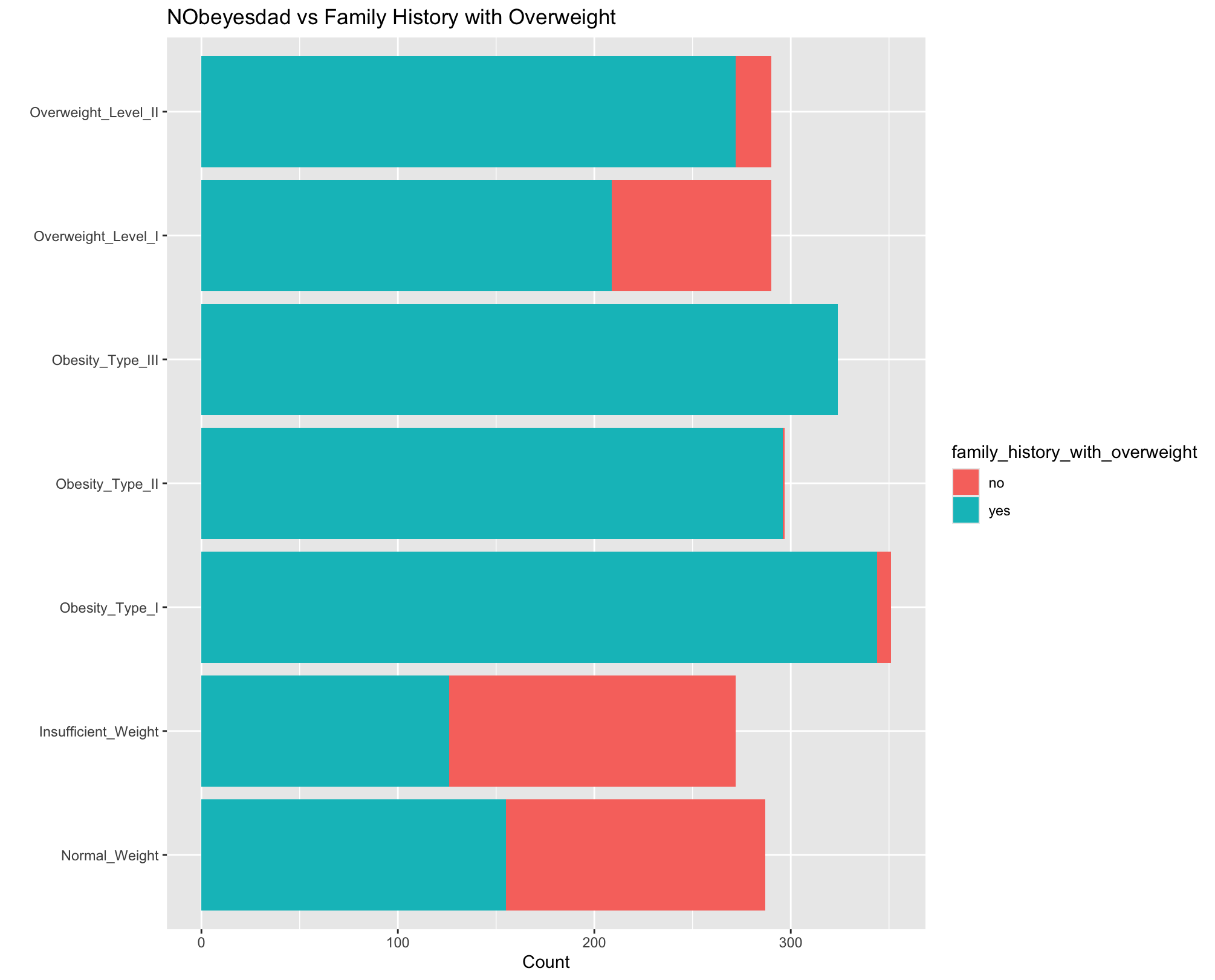
**Figure 3** Obesity Level Distribution

1. **NObeyesdad vs Age:** The box plot below shows that the age for “Normal\_Weight” is generally lower than Overweight and Obesity levels. This suggests that older people may be more prone to overweight issues. Additionally, there are outliers in each level but no removal is done to preserve the dataset's completeness for model training purposes allowing a more comprehensive assessment of model performance, especially in the real world context.

# 

**Figure 4** Box Plot for NObeyesdad vs Age

1. **NObeyesdad vs Family History with Overweight:** The stacked bar plot shows that in all the Obesity levels, a large majority of people have a family history with being obese. This may suggest a potential genetic influence on obesity.



**Figure 5** Stacked Bar Graph for NObeyesdad vs Family History with Overweight

Please refer to **Appendix A** for additional exploratory plots.

## **Predictive Models**

The 70/30 test split was adopted in our analysis as it is a common approach in machine learning to evaluate the performance of a model.

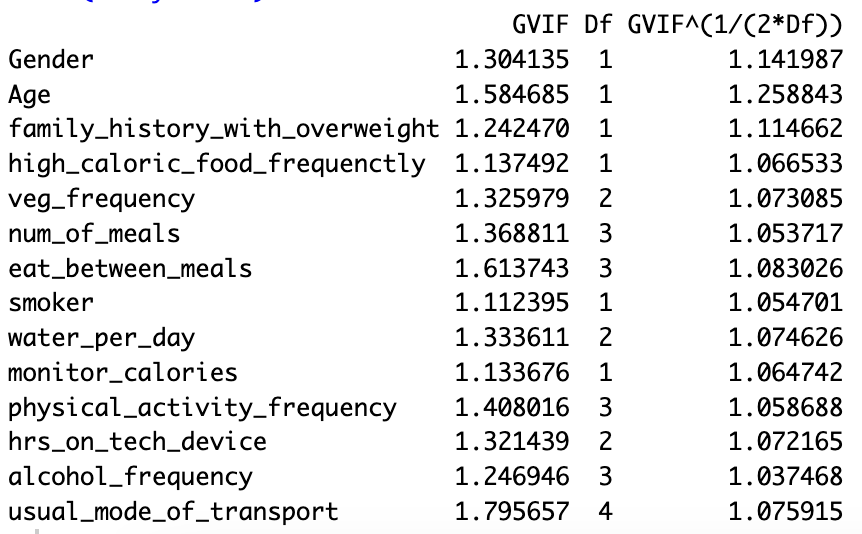
### **Multinomial Logistic Regression (MLR)**

*Methodology*

Unlike the simple logistic regression which only handles binary y-variable, MLR is able to handle multi-class classification. Similar to logistic regression, MLR requires independence among the data and no multicollinearity. (National University, 2024)

*Check for Multicollinearity*

As in R, the vif() function does not support MLR directly, so we have to work around it. One way is to create a dummy model with the same x-variables using glm() first and this allows us to apply the vif() function normally. (I. Rodriguez, 2023)

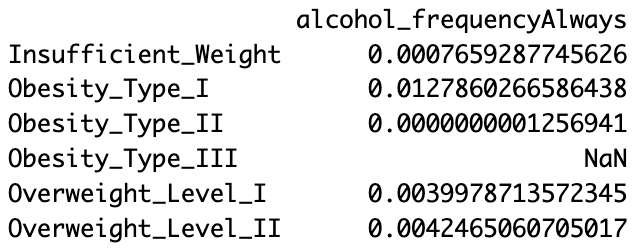


**Figure 6** Variance Inflation Factor (VIF) Values for All x-Variables

Based on the values in Figure 6 which are all below 2, we can conclude that there is no multicollinearity in our predictors.

*Warning Message*

After training the MLR, there is a warning message indicating that there are NaNs produced.

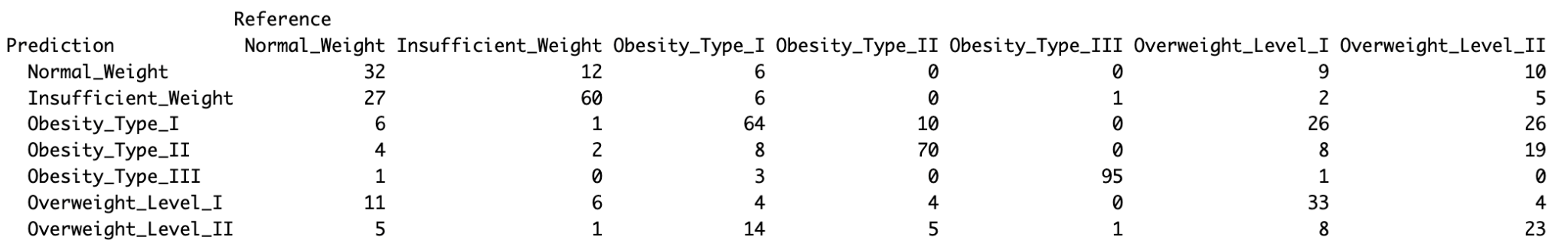


**Figure 7** Snippet of MLR Partial Summary

After examining the summary, it is noticed that there is only one NaN produced under “alchohol\_frequencyAlways” as shown in Figure 7. This warning typically occurs when there is multicollinearity and when there is not enough data for specific levels. As we have concluded that there is no multicollinearity, we went back to inspect the data count for “alcohol\_frequency” and we found that there is only one entry with “Always”. This highlights the importance of having sufficient data in all levels for effective modelling.

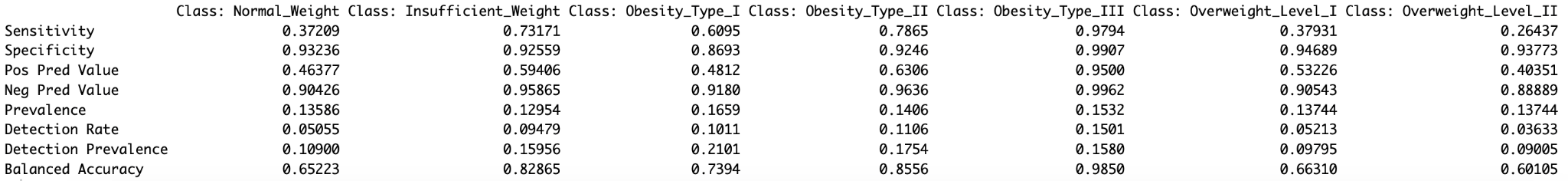
*Specifications*

Using the MLR trained, the accuracy obtained is 0.5955766 meaning that the model correctly predicted 59.56% of the cases.



**Figure 8** Confusion Matrix for MLR (Test Set)

According to Figure 8, the model performs well in predicting specific levels, such as “Obesity\_Type\_III”, which has a high number of true positives. However, it struggles with the other weight statuses, especially “Normal\_Weight”, “Overweight\_Level\_I” and “Overweight\_Level\_II”, showing a significant number of misclassifications.



**Figure 9** Statistics for MLR (Test Set)

Sensitivity also known as recall measures the ratio of true positive predictions to the sum of true positives and false negatives in the dataset. (Acharya, 2024) There are 3 levels, “Normal\_Weight”, “Overweight\_Level\_I” and “Overweight\_Level\_II” with exceptionally low sensitivity of below 0.4, suggesting that the model struggles to accurately predict in these categories.

“Pos Pred Value” in Figure 9 represents the precision score. A high precision score shows that the model is predicting little false positive cases. (Acharya, 2024) Most of the levels have positive predictive value ranging from 0.4 to 0.63 except “Obesity\_Level\_III” with 0.95 indicating the model prediction for “Obesity\_Level\_III” is likely to be accurate compared to the others.

### **CART**

*Methodology*

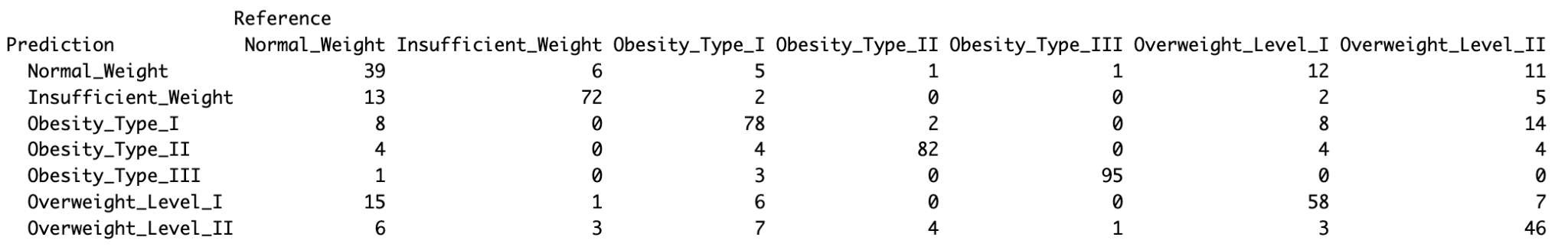
CART serves as a general term for decision trees and in our case, we are using classification trees where it will determine which weight status our target variable will fall into. CART outputs a tree-like structure where each branch is split into a weight status. CART will check through all criteria and choose the best in terms of reducing the impurity of the subsets for the splits. (GeeksforGeeks, 2024)

*Pruning of Tree*

Letting the tree grow to its maximum will cause overfitting of the data hence, it is important to prune the tree. Pruning often requires balance between cost complexity and the accuracy obtained. Using the 1SE rule, it was found that the optimal cp is 0.0007592671.

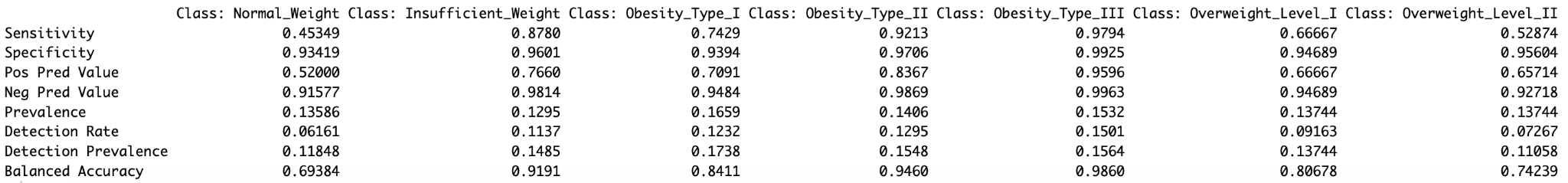
*Specifications*

Using the pruned tree obtained, the accuracy obtained is 0.7424961 meaning that the model correctly predicted 74.25% of the cases.



**Figure 10** Confusion Matrix for CART (Test Set)

Confusion matrix was plotted to see how well the model is predicting for the various weight status. Looking at Figure 10, we can notice that the model has more prediction errors in “Normal\_Weight” where it is wrongly classified into the other weight status.



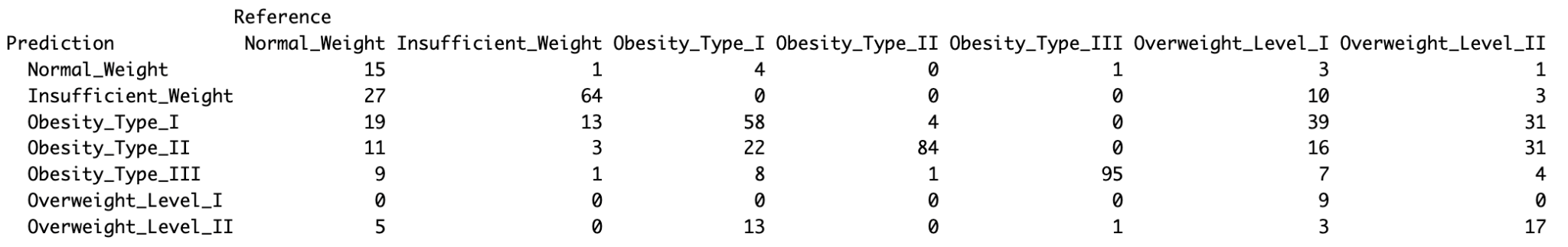
**Figure 11** Statistics for CART (Test Set)

In Figure 11, we notice that “Insufficient Weight”, “Obesity\_Type\_II” and “Obesity\_Type\_III” have very high sensitivity compared to “Normal\_Weight” indicating that the model is not effective in predicting “Normal\_Weight” accurately most of the time.

We can see that “Obesity\_Type\_II” and “Obesity\_Type\_III” have a high positive predictive value of above 0.8, indicating more accurate classification compared to “Normal\_Weight”

*Further Pruning*

It was noticed that the optimal tree is too huge. Attempt was made to further prune the tree down to where it contains ten terminal nodes (cp = 0.01288515) but results were undesirable.



**Figure 12** Confusion Matrix for CART (cp = 0.01288515)

From Figure 12, we can see that there are many wrong classifications indicating a poor model. Also, the accuracy for this tree dropped to 0.5402844 which is even lower than the MLR. Considering the number of x-variables we have, the multi category y-variable and the importance of accuracy in our context, we decided to not prune it further and use the optimal tree obtained using the 1SE rule.

### **Random Forest**

*Methodology*

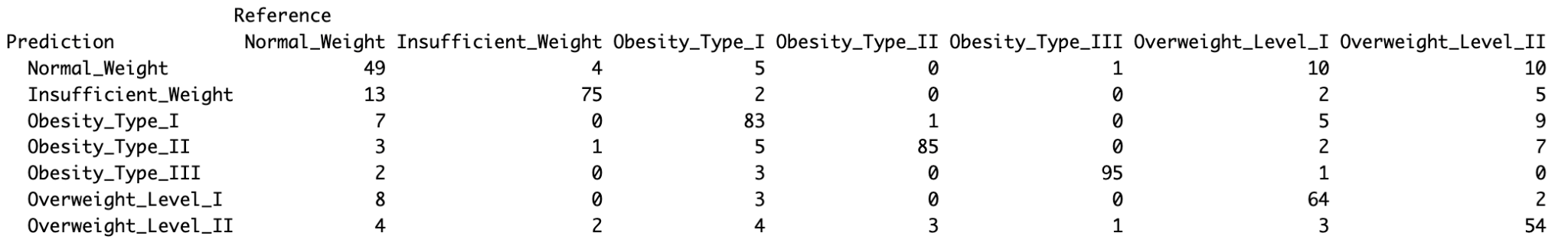
Random Forest generates and combines the results of multiple decision trees into one single result. It is versatile, making it suitable for classification tasks. Unlike decision trees which tend to suffer from overfitting, the Random Forest model does not face such a problem. However, as a trade off, Random Forest often requires more time and computational resources to generate. (Sruthi, 2024)

There are two additional parameters to note when generating a Random Forest in RStudio:

* ntree - This controls the number of trees generated. Generally more trees lead to a better performance but it will also cause longer computational time and higher computational cost. We used the typical value of 500 for our analysis.
* mtry - This specifies the number of features to be considered at its split. We used the default value which is the square root of the total number of all predictors. (Bhalla, n.d.)

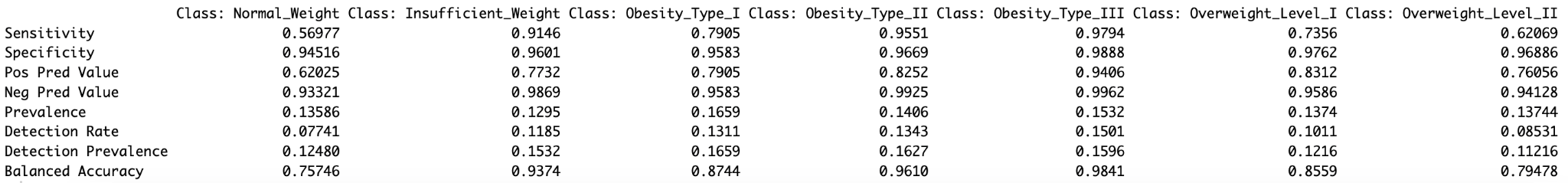
*Specifications*

Using the Random Forest generated, the accuracy obtained is 0.7977883 meaning that the model correctly predicted 79.78% of the cases.



**Figure 13** Confusion Matrix for Random Forest (Test Set)

According to the confusion matrix in Figure 13, the model demonstrated a decent level of accuracy, especially in cases of “Obesity\_Type\_III” and “Obesity\_Type\_II”. However, there are some cases where there is misclassification, especially between “Normal\_Weight” and “Insufficient\_Weight” and also among the other weight statuses.



**Figure 14** Statistics for Random Forest (Test Set)

Sensitivity is highest for “Obesity\_Type\_II” and “Obesity\_Type\_III”, suggesting the model is effective at predicting these cases most of the time. However, sensitivity is low for “Normal\_Weight” and “Overweight\_Level\_II” which indicates a higher rate of inaccurate prediction in these levels.

The high positive predictive value for “Obesity\_Type\_III” and “Overweight\_Level\_I” suggests that these levels have very accurate predictions. However, the low positive predictive value for “Normal\_Weight” indicates that there are more misclassifications compared to other weight status.

## **Best Model Selection**

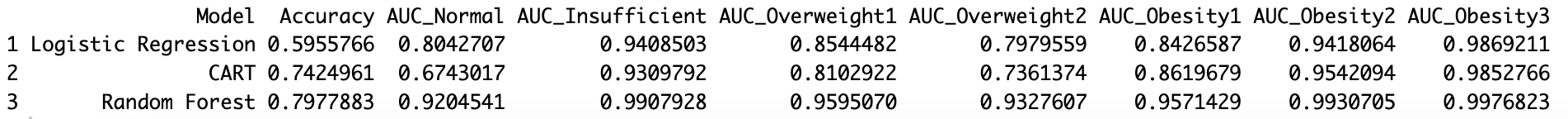
*ROC Curve & AUC*

Receiver Operating Characteristic (ROC) curves are useful for evaluating classification models, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) across various thresholds. (Classification: ROC and AUC, n.d.) This visualisation is important in our context where minimising false negatives is more important than controlling false positives as failure to identify individuals who are actually obese would prevent them from getting the necessary attention.

ROC curves generate a single value which is area under curve (AUC) that summarises the overall performance across all thresholds. The AUC summarises a model's performance across all thresholds. AUC of maximum 1.0 indicates perfect differentiation between classes. (Hajian-Tilaki, 2013) This value helps to make easy comparisons between different models, aiding the selection of the best model.

*Best Model*

Accuracy is the most common and used indicator in deciding a model effectiveness in making correct predictions. The Random Forest model achieved a high accuracy of 79.78% outperforming both MLR and CART that had the accuracy of 59.56% and 74.25% respectively. This difference highlights the Random Forest model’s strength in providing a more accurate classification across the different weight status.



**Figure 15** Accuracy Comparison between Models

When comparing AUC, Random Forest continues to perform better over the other two models. It has a consistent value of above 0.9 across all weight status compared to the other two models, where they have weaker classification in certain weight status. This suggests that Random Forest may be more effective for broader classification tasks.

Looking at the model itself, they all have their pros and cons. MLR is known for its simplicity but is unable to handle complex relationships. CART strives better when it comes to non-linear relationships but faces the risk of overfitting when not pruned properly. Random Forest has the strengths of CART without the risk of overfitting but its complexity often requires more computational resources.

Overall, based on a thorough comparison of accuracy, AUC values and the models’ strengths, Random Forest seems to be the best model for this analysis. Its highest overall accuracy and robust performance across all the weight statuses suggests that it is well suited for our context.

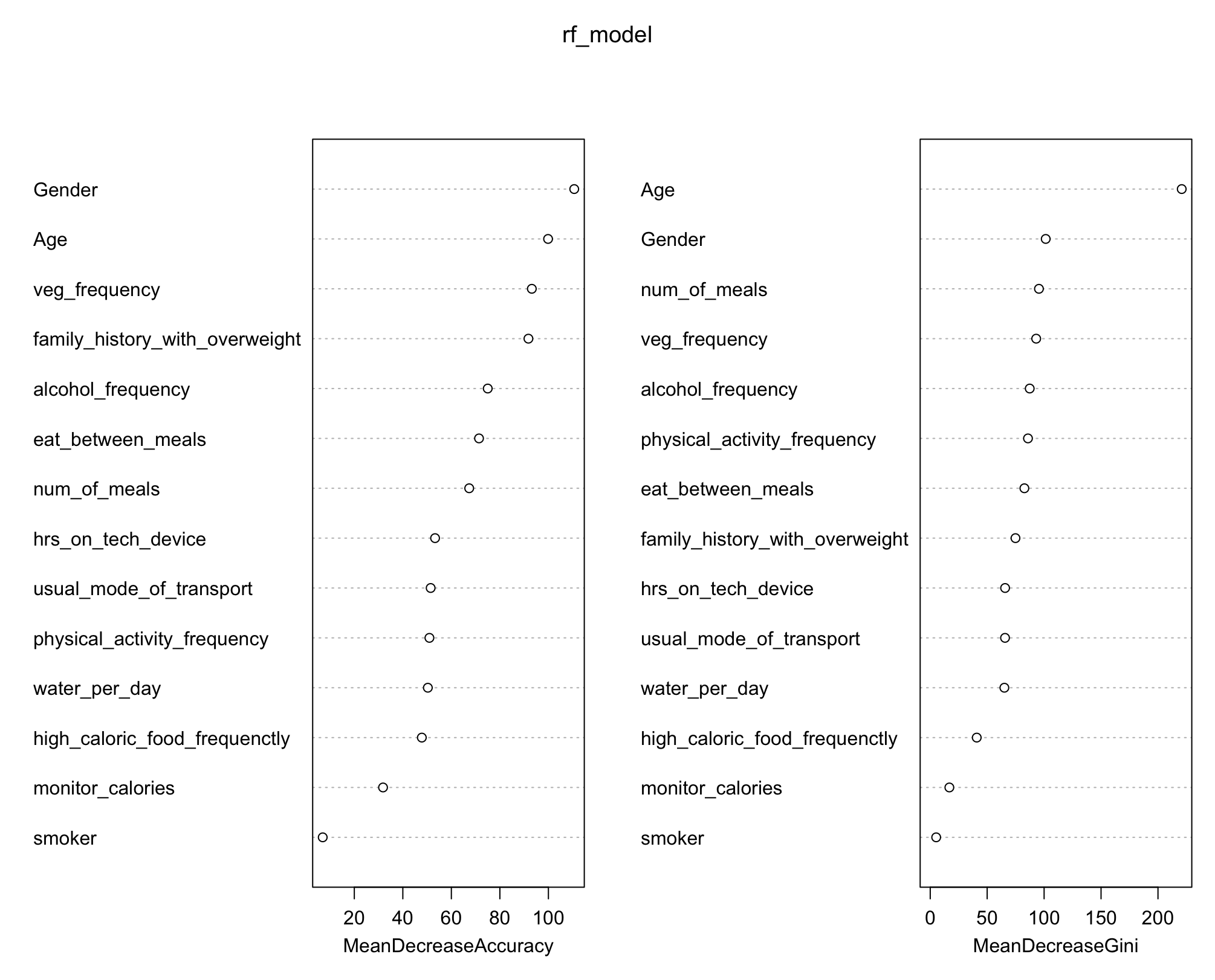
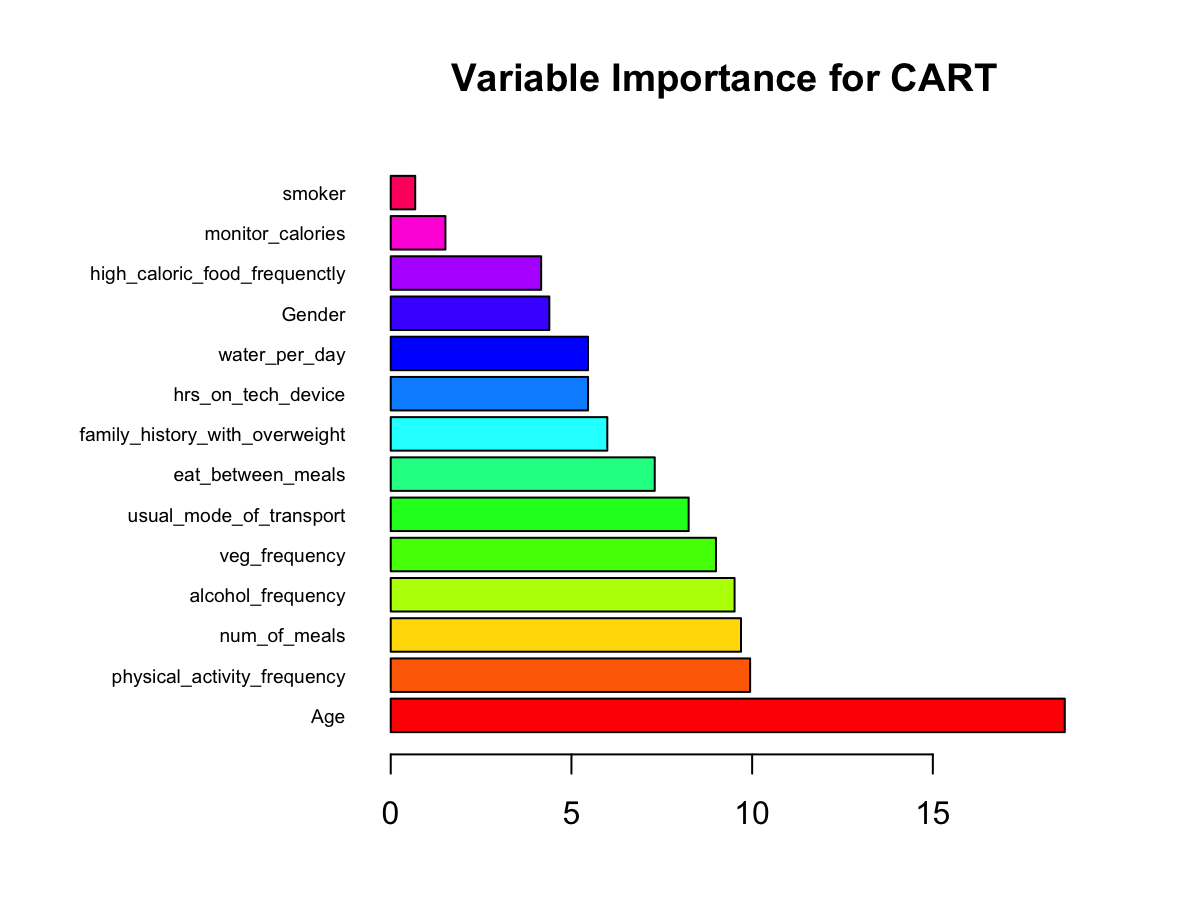
# **Analysis & Insights**

## **Models Findings & Insights**

The predictive model analysis has yielded some significant insights that can guide effective public healthcare strategies. From our analysis, Random Forest trees appear to be the most accurate with the best accuracy rate of 79.78% in predicting an individual’s weight class with respect to multiple predictor variables. The high accuracy rate is beneficial for key stakeholders since it allows them to make data-driven decisions such as health interventions and resource allocation. Furthermore, the Random Forest tree model’s ability to come up with reliable predictions can help health organisations identify specific groups with the highest risks more effectively and come up with targeted outreach and prevention programs. All this is such that they could combat the increasing trend of obesity around the world.

## **Variable Importance & Statistical Significance**

Variable importance is an indicator of the relative influence of a variable when it comes to the construction of decision trees. It measures how often a variable is used as a split criteria during the tree building process and how much mean squared error as well as the node purity improved as a result. Figure 13 shows the variable importance of the predictor variables for CART Model whereas Figure 14 shows the importance for RandomForest Model.



**Figure 16** Variable importance for CART **Figure 17** Variable importance for RandomForest

After running our CART and RandomForest models, it can be observed that in the CART Model, the top variables are “Age”, “Physical Activity Frequency" and “Number of Meals” whereas for Random Forest, the top variables are “Age” and “Gender”. Therefore, when comparing both models it seems that the “Age” variable is the most important variable.

Confidence intervals is also another benchmark that could be used to evaluate the results of the multinomial logistic regression model where it provides valuable insights into the reliability of the estimated odd ratios (OR) for each variable. The model gives a range of values which we can expect the true population parameter to fall on, with a specified confidence level of 95%. For instance in **Appendix D**, we can take a look at the confidence interval of the “Age” variable. With a reference level of Normal Weight, the “Age” variable in other weight classes is statistically significant with confidence intervals ranging 0.73 to 0.88 in Insufficient Weight class and 1.03 to 1.16 in Overweight Level 1 class. These 2 examples showed that as long as the interval ranges do not include 1, it suggests that the variable has a statistically significant association. However, for Obesity Type 3, “Age” is not a significant variable with reference to the Normal Weight class which might suggest that there are more important factors at play that would lead an individual to undergo a drastic weight class change.

## **Proposed Enhancement/Solutions**

By analysing the data from the dataset, we can observe the trends and factors affecting obesity. Leveraging AI to automate this analysis and facilitate consistent monitoring of these trends allows for more effective design and implementation of targeted programs but aids relevant organisations in tracking the effectiveness and reach of these programs. Our proposed solution is thus a targeted obesity and intervention system in the form of an app. When first downloaded, users will input information like age, gender and dietary habits into the app which uses these inputs as variables to predict and categorise users into different obesity risk groups. Random Forest would be the model used for this prediction due to the high accuracy as shown in our analysis. Based on the different risk levels, high, moderate and low, the government and health organisations can tailor programs of different intensities that are personalised for individual needs. For example, programs for high-risk groups could include customised diet and exercise plans, mandatory participation in health programs like nutrition counselling and national physical activity events, and regular medical checkups. Users can track the completion of programs and plans to monitor their progress, and these results can be used to update and modify the programs as necessary. Moderate-risk groups may only require dietary guidelines and notifications that encourage users to incorporate fruits and vegetables into their consumption patterns and participate in physical activity programs. Unlike high-risk groups, moderate-risk users will have programs targeted at promoting healthy habits instead of strict regimes. For low-risk groups, general campaigns, and online resources that have nutrition tips and self-guided wellness plans may be sufficient.

Examining our model results, age, alcohol and vegetable consumption frequency are significant factors of obesity. The app uses these results to appeal to users, such as in the form of interactive digital design and in-app games for younger users to encourage the use of the app and participation in the challenges. For older users, a rewards program could be used where points could be redeemed for rewards like shopping vouchers. To target alcohol use, the app could ask questions to derive the factors behind the frequency of alcohol use and suggest mitigating steps. For example, for alcohol use due to stress or other mental health reasons, the app may suggest online resources or stress-relieving activities like breathing exercises as alternative coping mechanisms. Similarly, the app can learn about users' consumption preferences through questions and consumption patterns through machine learning, then suggest methods to include vegetables in meals in ways that appeal to users to increase vegetable consumption.

As users engage more with the app, suggestions for individuals could be refined to better target individuals, creating a positive cycle where users are more willing to complete tasks while reducing their obesity risk. Relevant organisations are also able to use this data to better design campaigns that appeal more to individuals and reach a wider range of individuals on a national level to target the issue of obesity.

# **Evaluation**

## **Effectiveness Assessment**

It is an optimal choice to select the Random Forest model because of its high predictive accuracy. It is capable of analysing trends using a wide range of variables like age, alcohol consumption and dietary habits to accurately predict obesity risk levels. This ensures that the interventions are highly targeted towards their unique needs and risk levels. The high accuracy of Random Forest reduces the likelihood of misclassification, lowering the chances of grouping a high-risk individual in the low-risk group for example. This is crucial for individuals in the high-risk category who need intensive support, preventing wrong interventions. The incorporation of interactive designs, games and rewards in the mobile application motivates and incentivises regular user interaction, which builds positive reinforcement and supports long-term behavioural change. Due to the varying categorisation of users, it ensures that appropriate resources are strategically allocated based on their needs and this increases their commitment to the programme, which makes a more significant impact in their health levels. The app has a greater impact on user health since it optimises health outcomes while sparing low-risk users from needless actions. The organisation that created the mobile application can continuously make improvements to the design, features and programmes that are based on users’ behaviours and preferences by utilising the user data. This may reach more users on a national level.

## **Limitations**

Random Forest models are computationally demanding when dealing with large datasets. This could potentially increase operational cost as the maintenance of the app performance requires higher processing power while the number of users increases. The health situation of individuals in the high-risk group is more dire, hence more resources will be allocated to them. This may strain resources if the number of users increases steadily without sufficient resources and support from the relevant organisations. As a result, it may affect the quality and accessibility of services for all users. As users have to input personal information, it may erode users trust due to the rise of recent data breaches globally. It is of paramount concern to regulate data privacy and mandate security measures to handle and protect sensitive personal information. Due to underreporting or conforming to social norms, the self-reported data may be inaccurate and it might impact the model’s accuracy and the effectiveness of the programmes. Additionally, it is possible that many people who battle obesity are unaware of the seriousness of the condition or its potential health risks. The lack of understanding and denial hinders them from downloading the app. The effectiveness of the app may be undermined due to the limitations stated above.

## **Sustainability**

As Random Forest's accuracy guarantees that suggestions change in response to user behaviour, it serves as a basis for the app's long-term development. A feedback loop that improves predicted accuracy over time is created when initial classifications are accurate because data gathered from continuing encounters will more successfully update the model. The software can enhance its interventions and recommendations as it collects more data, initiating a self-sustaining cycle of more targeted obesity control. Certain features of the gamified elements, rewards programme and customised content can be modified to suit the user’s preference more closely as they continuously use the app. To ensure the continuity of the app, we can collaborate with the Ministry of Health and seek support from the Lively Places Fund (SupportGoWhere, n.d.) to receive sustainable funding, resources and expert opinion. By helping users who move to lower risk levels and minimising the need for expensive interventions, the tiered approach enables the app to stay relevant as users advance, promoting long-term sustainability in the management of obesity.

# 

# **Conclusion**

With obesity rates rising across all age groups, Singapore is facing similar challenges to the growing worldwide health crisis of obesity. This report suggests using machine learning, such as Logistic Regression, CART and Random Forest, to predict obesity risk levels based on lifestyle factors such as diet, physical activity, and alcohol intake. Through the identification of high-risk individuals, the model facilitates targeted health interventions. The proposed solution is a mobile app that categorises users by obesity risk and tailors programs accordingly. Long-term participation is encouraged by features like workout recommendations, food programs, and rewards. The model’s predictive accuracy is used to produce precise, data-driven strategies for reducing obesity, while partnerships with relevant authorities and organisations can help maintain the app’s viability.

## 

# **References**

Acharya, N. (2024, February 15). *Understanding precision, recall, F1-score, and support in Machine Learning Evaluation*. Medium. https://medium.com/@nirajan.acharya777/understanding-precision-recall-f1-score-and-support-in-machine-learning-evaluation-7ec935e8512e

Bhalla, D. (n.d.). *Random Forest in R: A step-by-step guide*. ListenData. https://www.listendata.com/2014/11/random-forest-with-r.html

DataKulture. (n.d.). *Why is machine learning forecasting better than traditional methods?* DataKulture. https://datakulture.com/blog/why-is-machine-learning-forecasting-better-in-traditional-methods/#:~:text=ML%20forecasting%20models%20have%20always

GeeksforGeeks. (2024, September 19). *CART (Classification And Regression Tree) in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/

Google. (n.d.). *Classification: ROC and AUC*. Google Machine Learning. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Hajian-Tilaki, K. (2013). *Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation*. U.S. National Library of Medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC3755824/

HealthHub. (2022, November 15). *Obesity in Singapore: It’s Not a Small World*. HealthHub. https://www.healthhub.sg/live-healthy/its-not-a-small-world-after-all

How have costs associated with obesity changed over time? (n.d.). *Peterson-KFF Health System Tracker*. https://www.healthsystemtracker.org/chart-collection/how-have-costs-associated-with-obesity-changed-over-time/

IBM. (n.d.). *What is supervised learning?*. IBM. https://www.ibm.com/topics/supervised-learning

I. Rodriguez, C. (2023, October 8). *Multinomial Logistic Regression*. Carlos I. Rodriguez. https://www.carlosivanrodriguez.com/guides/statistics/logistic-regression/multinomial-regression/

Knapton, M., Lewin, R. J., & Doherty, P. J. (2011). Long-term cardiovascular conditions: the role of primary care. *The British journal of general practice : the journal of the Royal College of General Practitioners*, *61*(592), 659–660. https://doi.org/10.3399/bjgp11X606492

Lobstein, T., Jackson-Leach, R., Moodie, M. L., Hall, K. D., Gortmaker, S. L., Swinburn, B. A., ... & McPherson, K. (2015, June 20). Child and adolescent obesity: part of a bigger picture. *The Lancet*, *385*(9986), 2510-2520. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)61746-3/abstract

Obesity Evidence Hub. (2024, June 14). *Obesity trends in adults globally*. Obesity Evidence Hub. https://www.obesityevidencehub.org.au/collections/trends/adults-global#cite2476

Obesity OPEN. (2023, March 15). *Open Singapore*. Obesity OPEN. https://obesityopen.org/open-singapore/

National University. (2024, October 30). *Libguides: Statistics Resources: Multinomial Logistic Regression*. National University Academic Success Center. https://resources.nu.edu/statsresources/Multinomiallogistic

Ng, Marie et al. (2014, August 30) *Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013*. The Lancet, Volume 384, Issue 9945, 766 - 781. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)60460-8/abstract?jobId=1138&record=14&sort=4&category=1

Sruthi. (2024, October 8). *Understanding Random Forest Algorithm With Examples*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

SupportGoWhere. (n.d.). https://supportgowhere.life.gov.sg/schemes/L4IhetvC/lively-places-fund

Swinburn, B. A., Sacks, G., Hall, K. D., McPherson, K., Finegood, D. T., Moodie, M. L., & Gortmaker, S. L. (2011, August 27). *The global obesity pandemic: shaped by global drivers and local environments.* The Lancet, 378(9793), 804-814 https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)60813-1/abstract?cc=y%3D

The World Bank. (2017). (publication). *An Overview of Links Between Obesity and Food Systems: Implications for the Food and Agriculture Global Practice Agenda*.

https://documents1.worldbank.org/curated/zh/222101499437276873/pdf/117200-REVISED-WP-Obesity-Overview-Web-PUBLIC-002.pdf

Wisneski, C. (2024, January 5). 7 reasons why machine learning forecasting is better than traditional methods. *Akkio*. https://www.akkio.com/post/5-reasons-why-machine-learning-forecasting-is-better-than-traditional-methods

World Health Organization. (2015, March 1). Overweight and obesity. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

# 

# **Dataset**

UCI Machine Learning Repository. (n.d.). *Estimation of obesity levels based on eating habits and physical condition Data Set*. https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

# 

# 

# 

# 

# 

# 

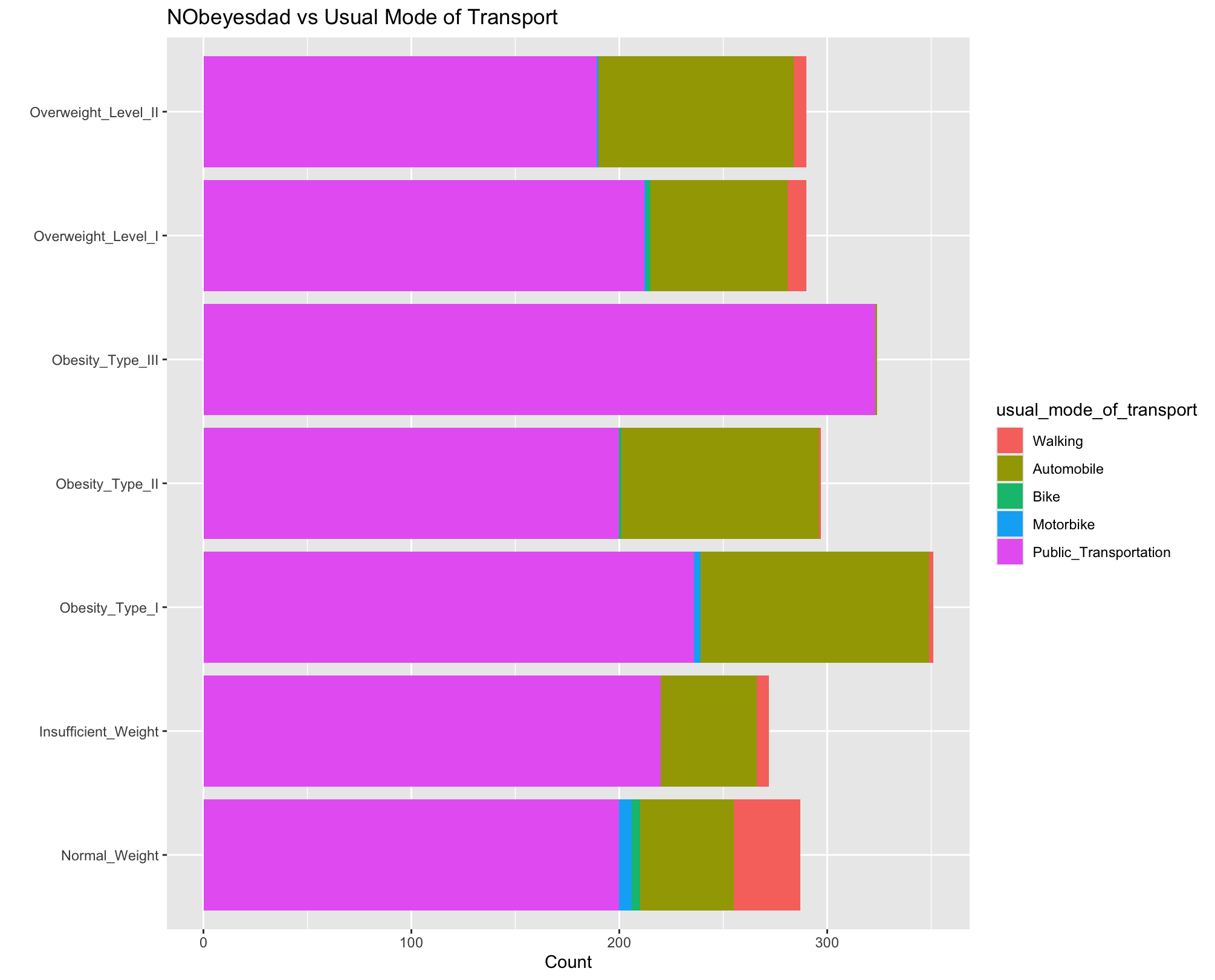
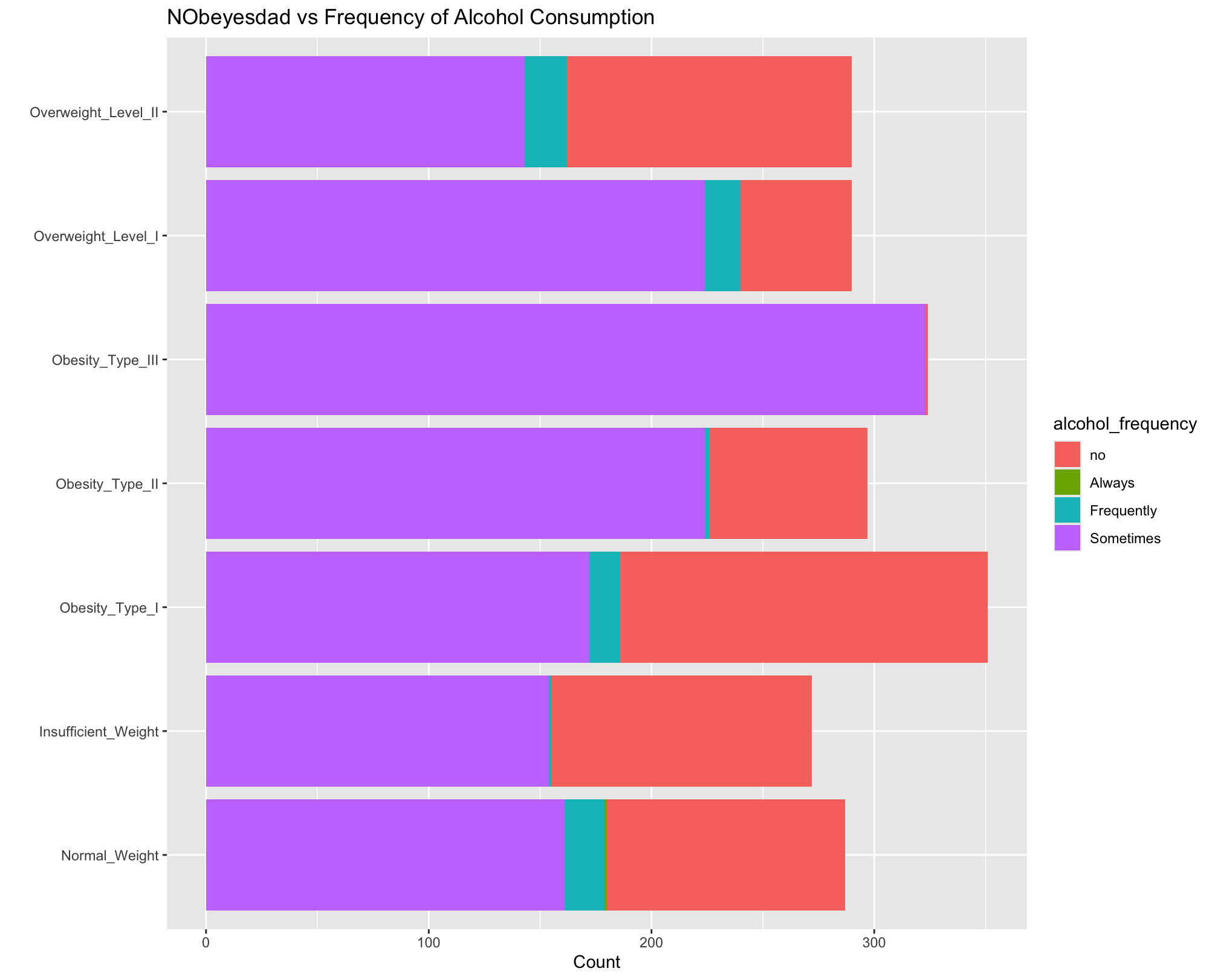
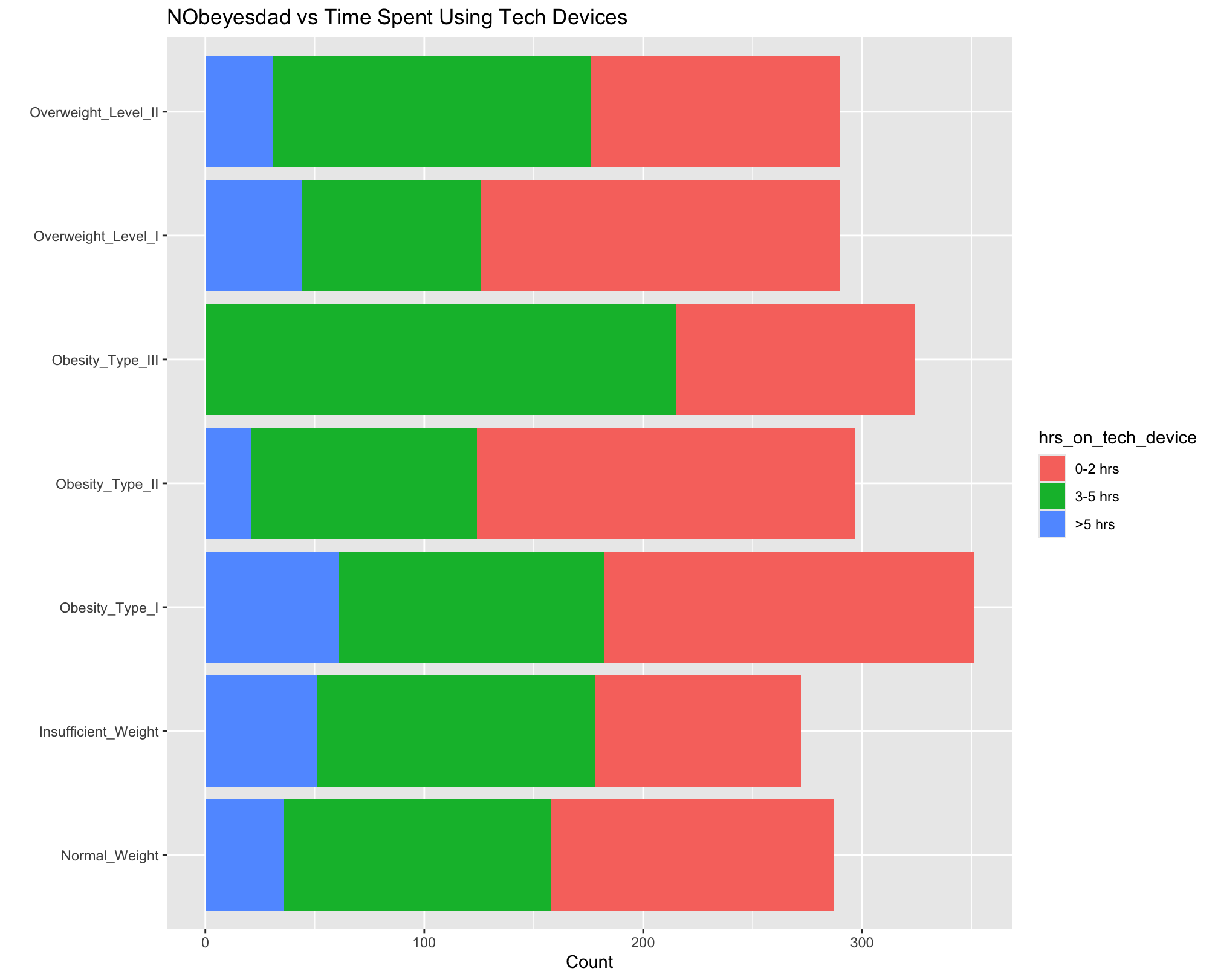
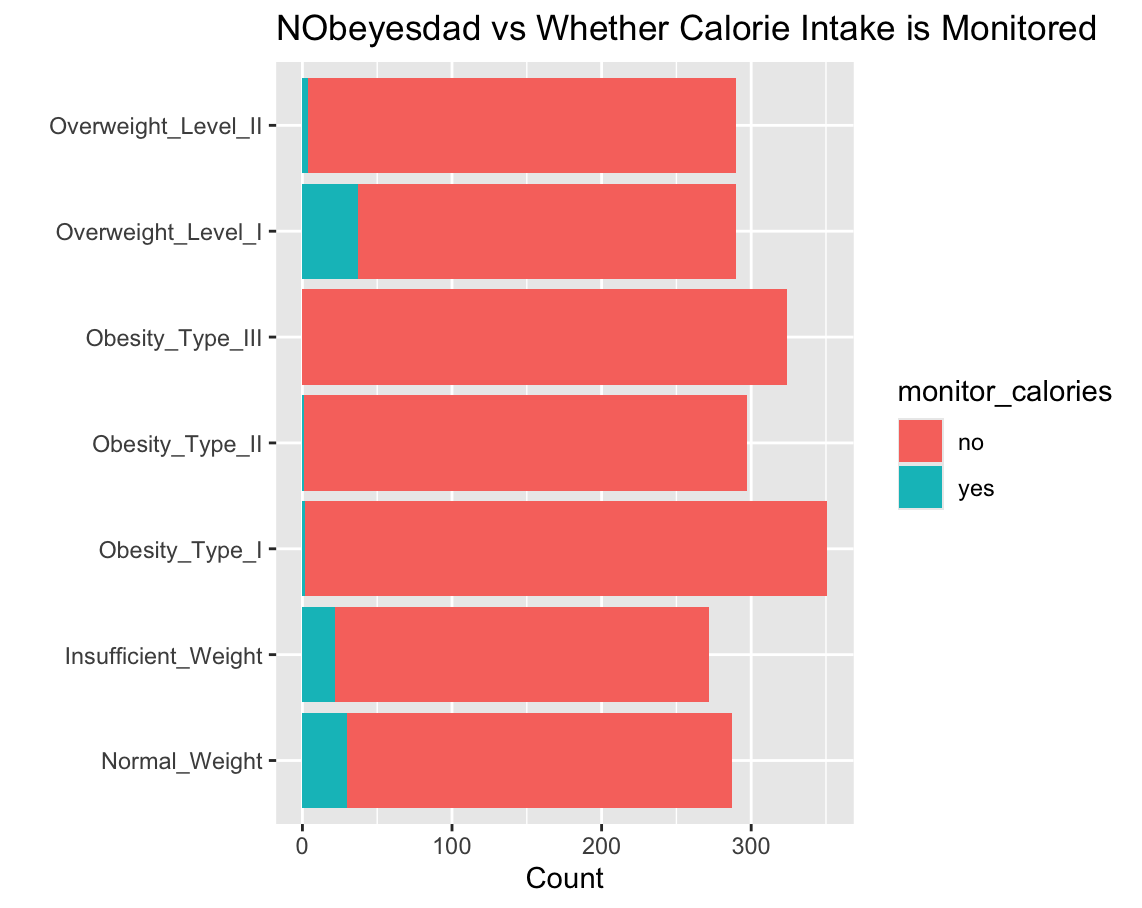
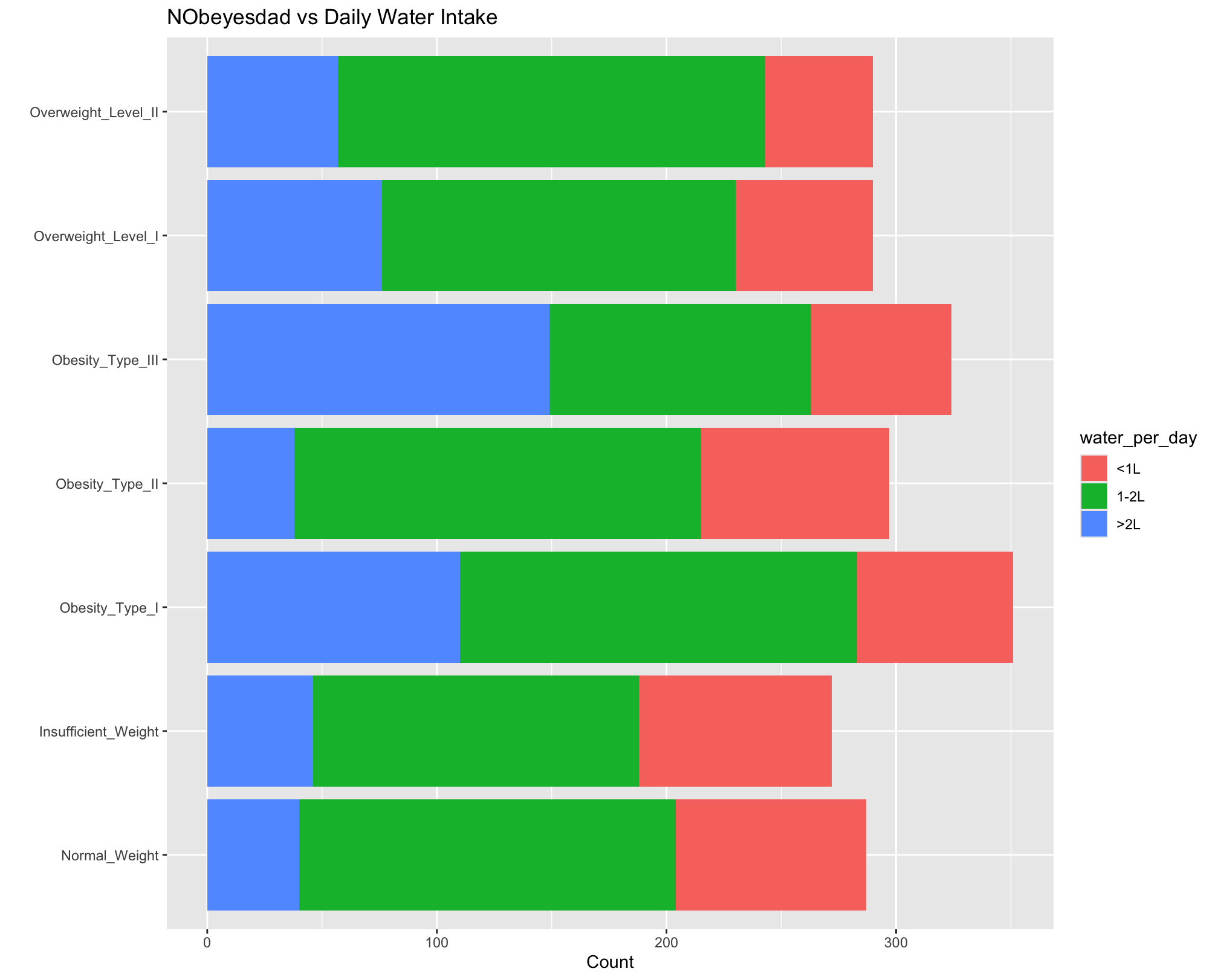
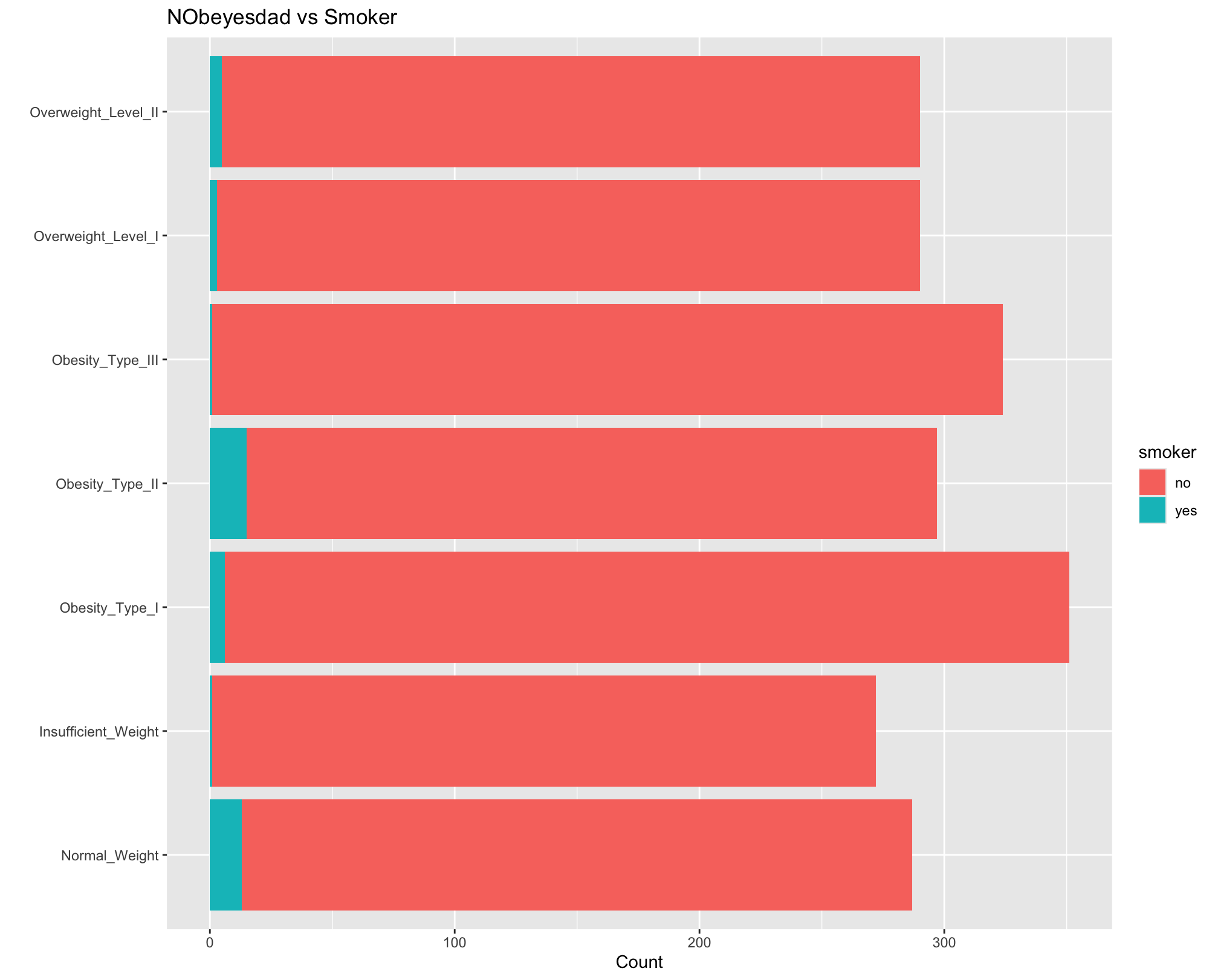
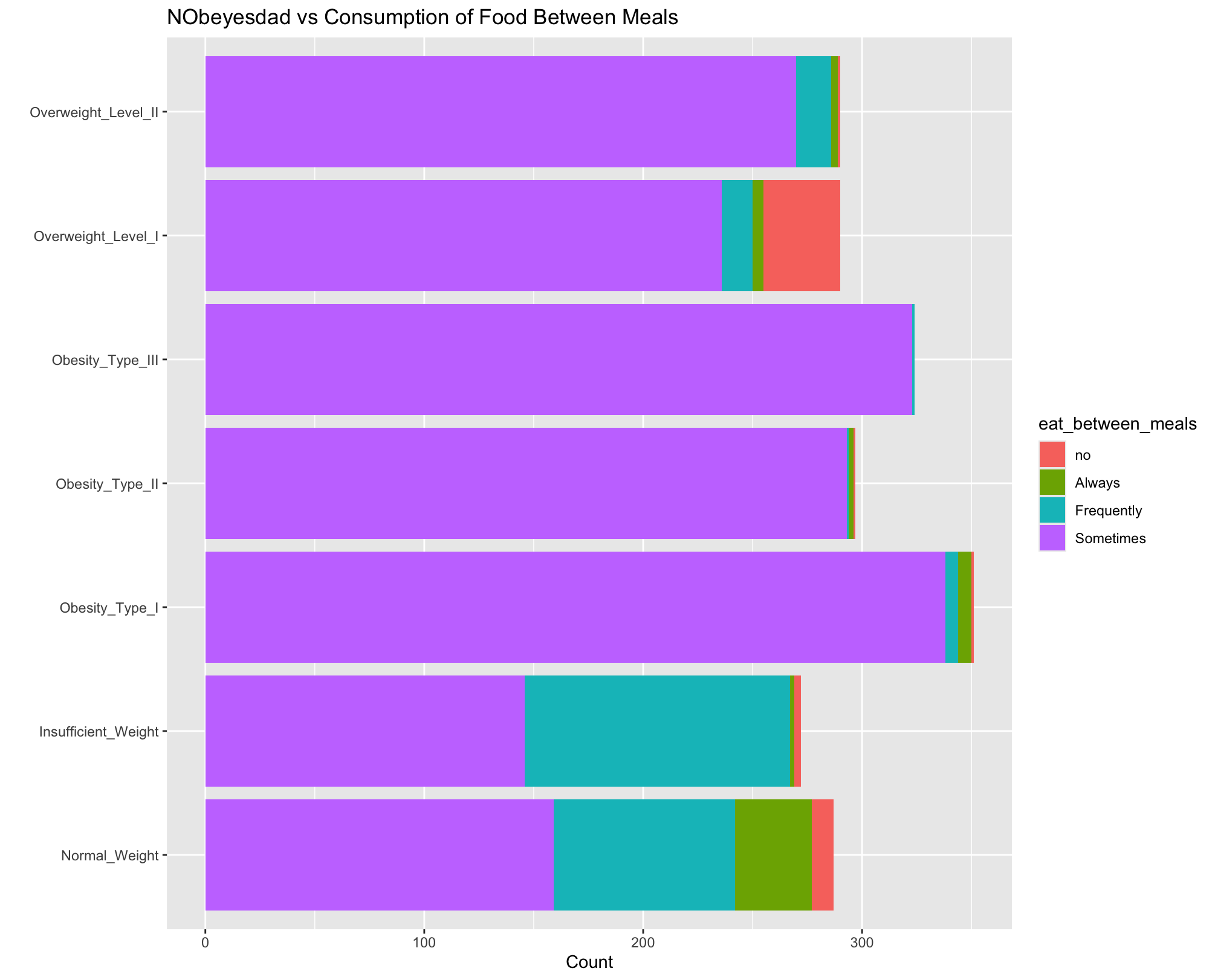
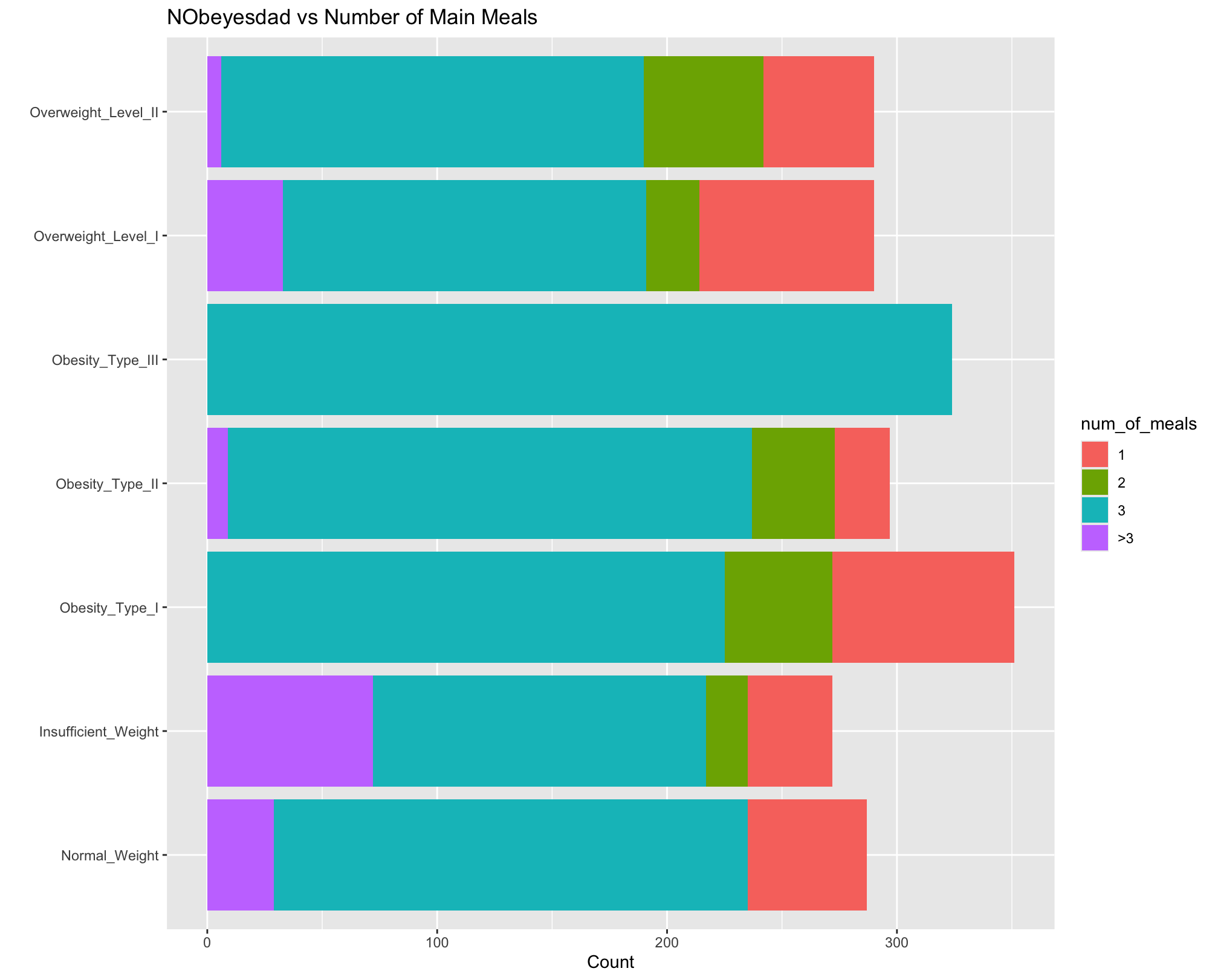
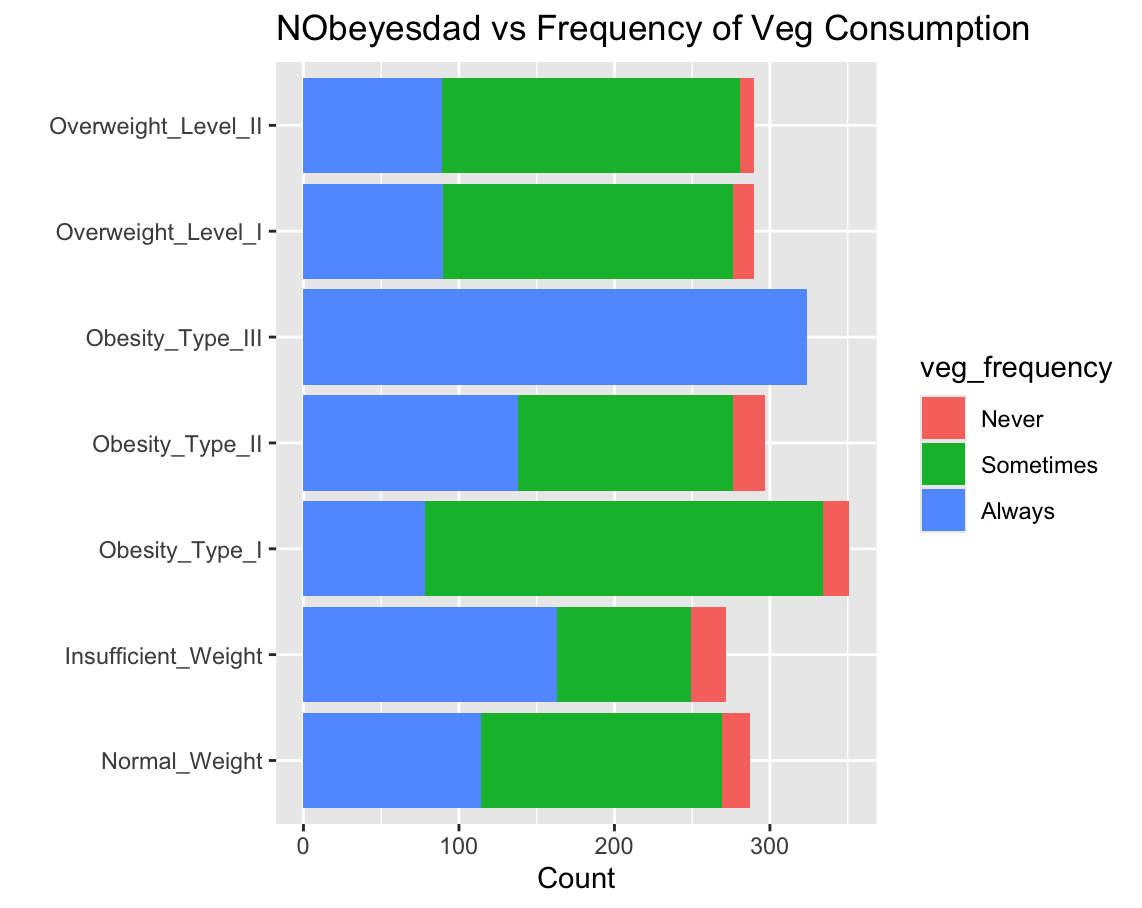
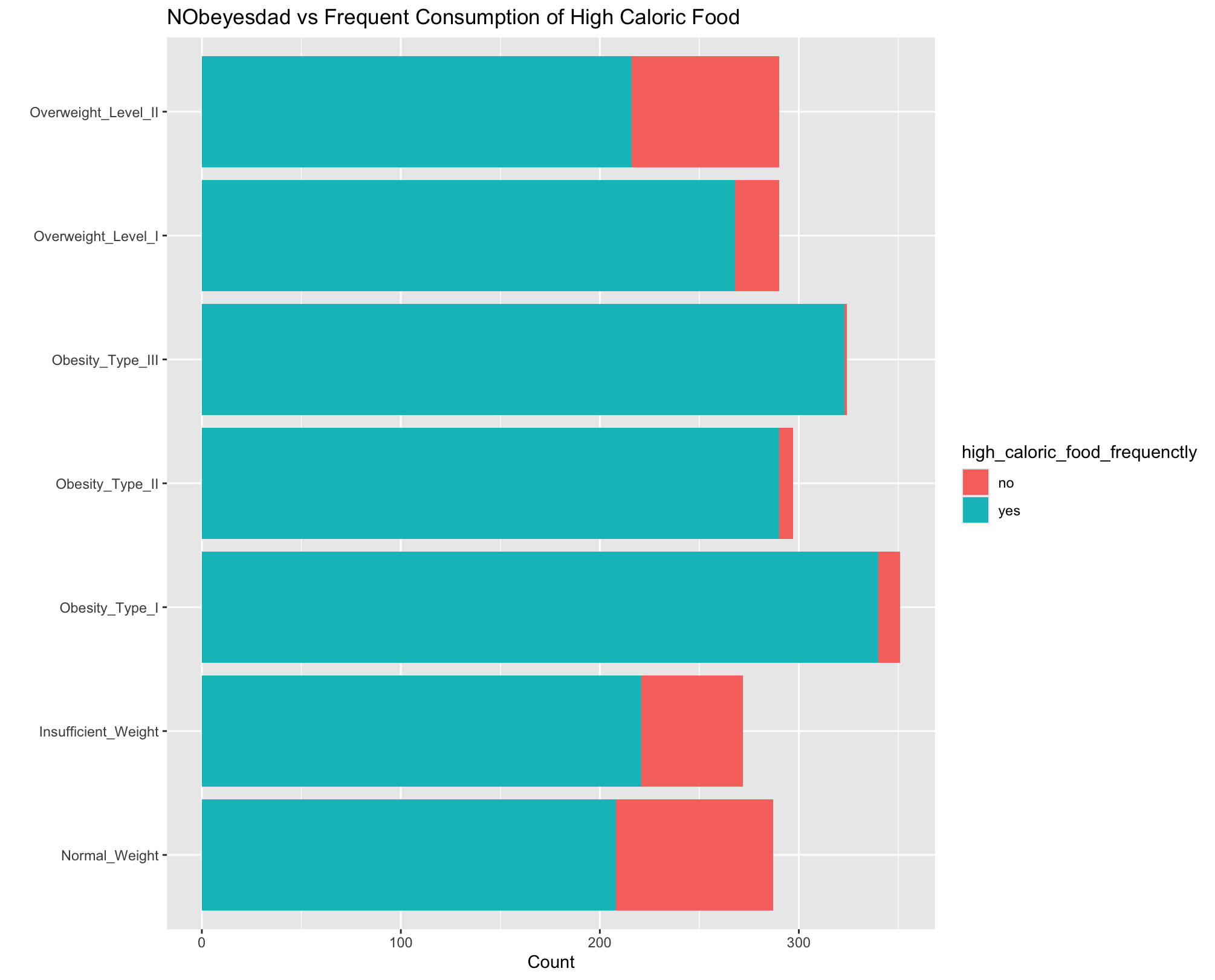
# 

# 

# 

# 

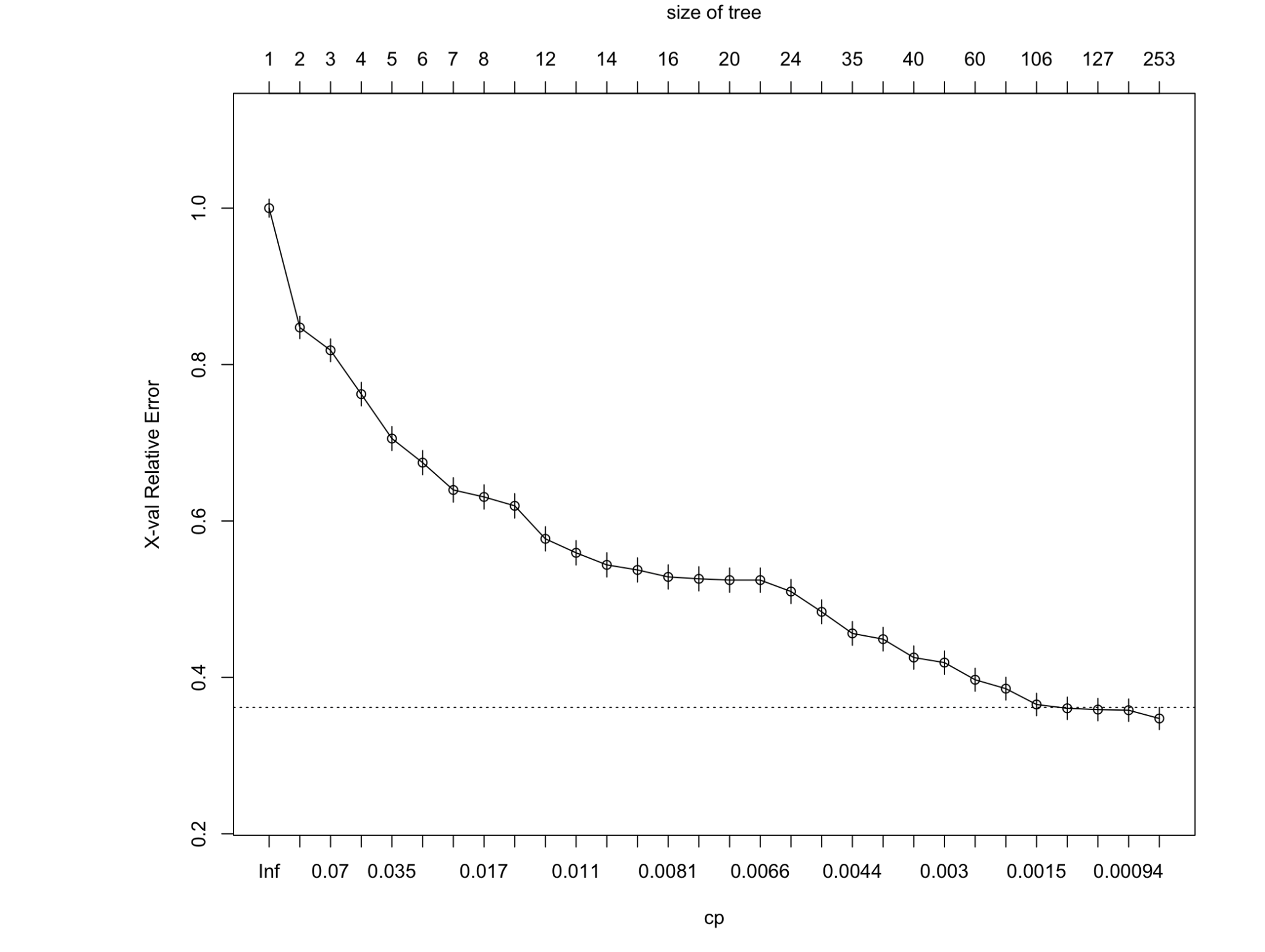
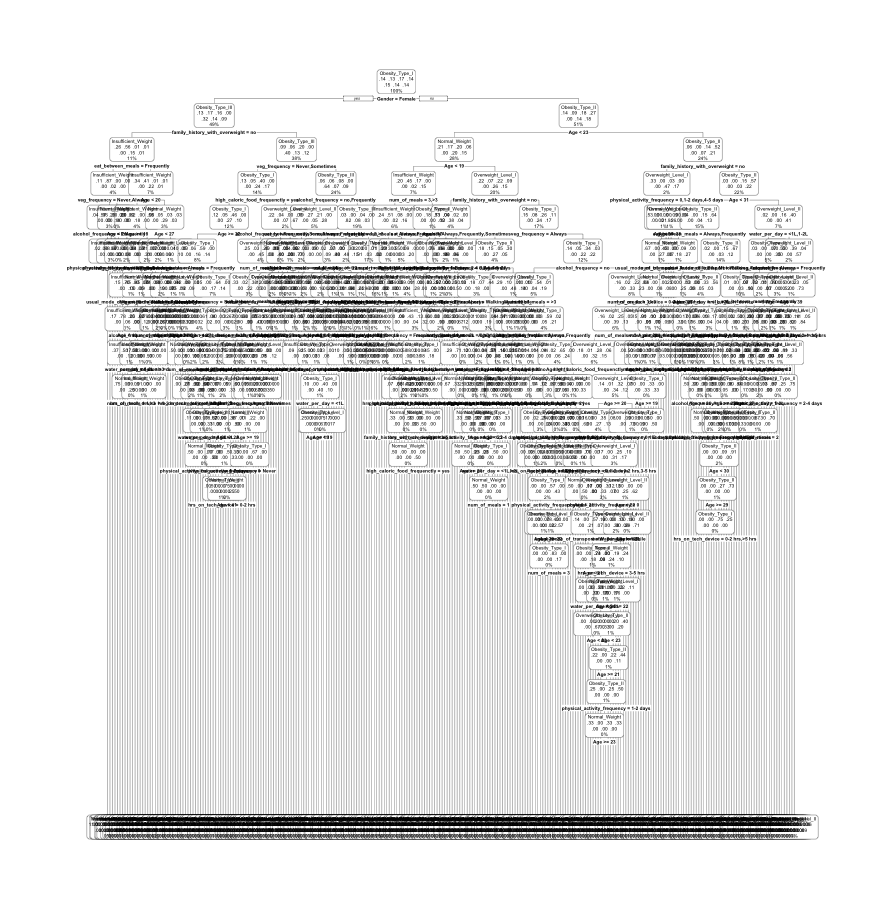
# **Appendix A (Exploratory Findings)**

​​

# 

# **Appendix B (CART Pruning)**

Optimal Tree



Alternative Tree After Further Pruning (cp = 0.01288515)

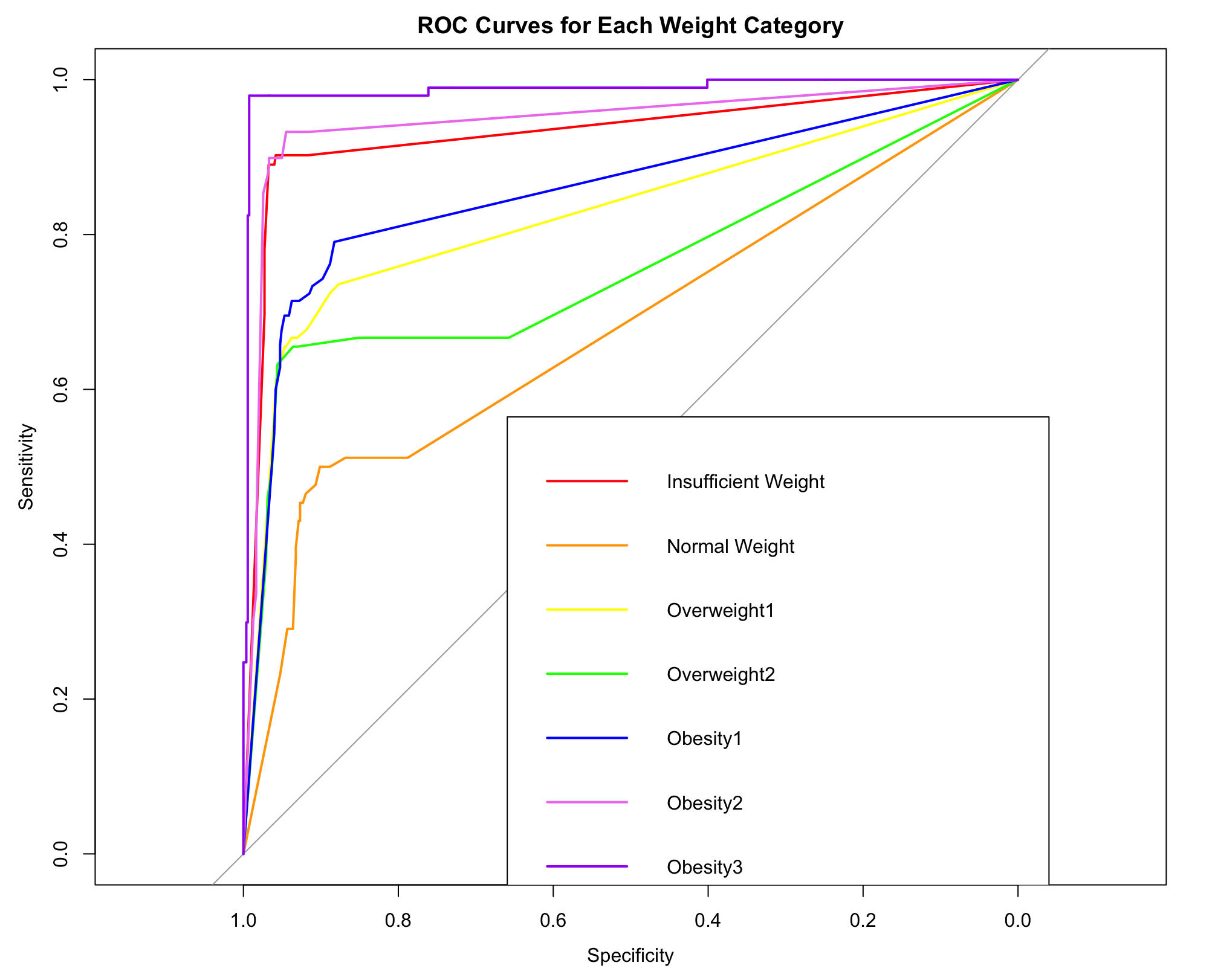
# 

# **Appendix C (ROC Curve Graph)**

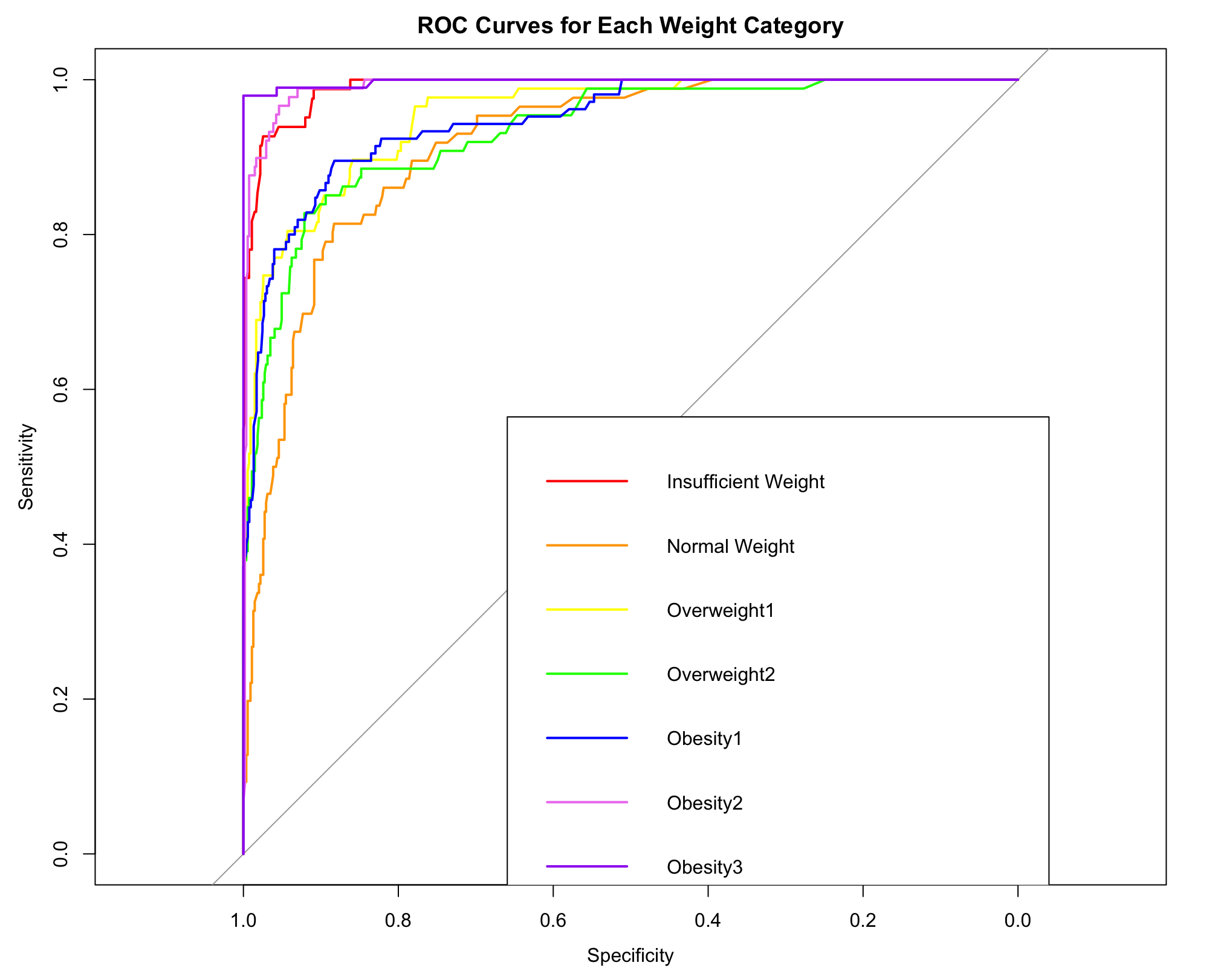
Multinomial Logistic Regression



CART



Random Forest



# **Appendix D (Confidence Interval)**

