Obesity Level Prediction using Decision Tree Models

ASDS 6303 – 2024 Fall

Lijuan Tang / Hariharan Selvam

Problem Statement

Obesity has become a global epidemic, posing significant health risks and economic burdens on individuals and healthcare systems alike. In recent decades, the prevalence of obesity has skyrocketed, driven by various factors, for example, lifestyles, unhealthy dietary habits, and inherent genetic tendencies. The consequences of obesity are far-reaching, including an increased risk of diseases like type 2 diabetes and certain types of cancer.

Addressing the obesity crisis requires a multifaceted approach, and one promising avenue is leveraging machine learning models to predict obesity levels based on individuals' health lifestyles and meal preferences. By harnessing the power of data-driven insights, these models can identify patterns and risk factors associated with obesity, enabling early intervention and personalized preventive strategies.

Moreover, these predictive models can be integrated into healthcare systems, empowering healthcare professionals to proactively identify at-risk individuals and implement targeted interventions. By addressing the obesity epidemic at its roots, we can mitigate the burden on healthcare systems, improve overall population health, and enhance the quality of life for countless individuals worldwide.

Dataset

This dataset contains information for estimating obesity levels among individuals in Mexico, Peru, and Colombia, focusing on their dietary patterns and physical fitness.

It has 2111 observations and 17 features, including the target label ‘NObeyesdad’. This is a multiclass question, the target label has 7 levels, including Insufficient Weight, Normal Weight,

Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

Here are the descriptions of the 16 explanatory variables:

Gender: Female/Male

Age: numeric value

Height: in meters

Weight: in kg

Family_history_with_overweight: Yes/No

FAVC: Do you eat high caloric food frequently - Yes/No

FCVC: Do you usually eat vegetables in your meals - Never/Sometimes/Always

NCP : How many main meals do you have daily - Between 1 to 2/ Three/ More than three

CAEC : Do you eat any food between meals? No/ Sometimes/ Frequently/ Always

SMOKE: Do you smoke? Yes/No

CH2O: How much water do you drink daily? Less than a liter /Between 1 and 2L /More than 2L

SCC: Do you monitor the calories you eat daily? Yes/No

FAF: How often do you have physical activity? I do not have /1 or 2 days/ 2 or 4 days/ 4 or 5 days

TUE: How much time do you use technological devices such as cell phone, videogames, television, computer, and others? 0–2 hours/ 3–5 hours/ More than 5 hours

CALC: How often do you drink alcohol? I do not drink/ Sometimes/ Frequently/ Always

MTRANS: Which transportation do you usually use? Automobile/Motorbike/Bike/Public Transportation/Walking

Explanatory Data Analysis

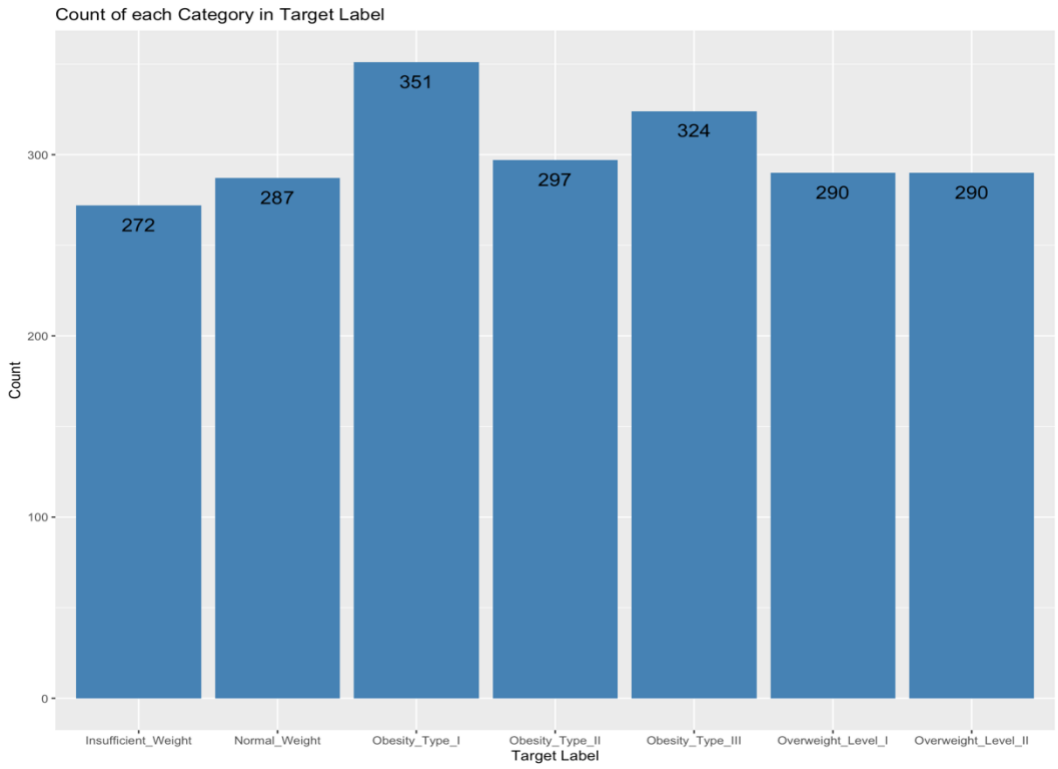
Target Label

The bar graph below illustrates the distribution of categories within the target label, obesity level.

On the x-axis are the various categories, including Insufficient Weight, Normal Weight, Obesity

Type I, Obesity Type II, Obesity Type III, Overweight Level I, and Overweight Level II. The y-axis displays the frequency of each category.

The highest count is observed in the normal weight category, totaling 351 instances, followed by obesity type I with 324 instances, and Obesity Type II with 297 instances. Both Overweight Level I and Overweight Level II categories have 290 instances each. Conversely, the insufficient Weight category exhibits the lowest count, with 272 instances. In general, the dataset demonstrates balance across various obesity levels.

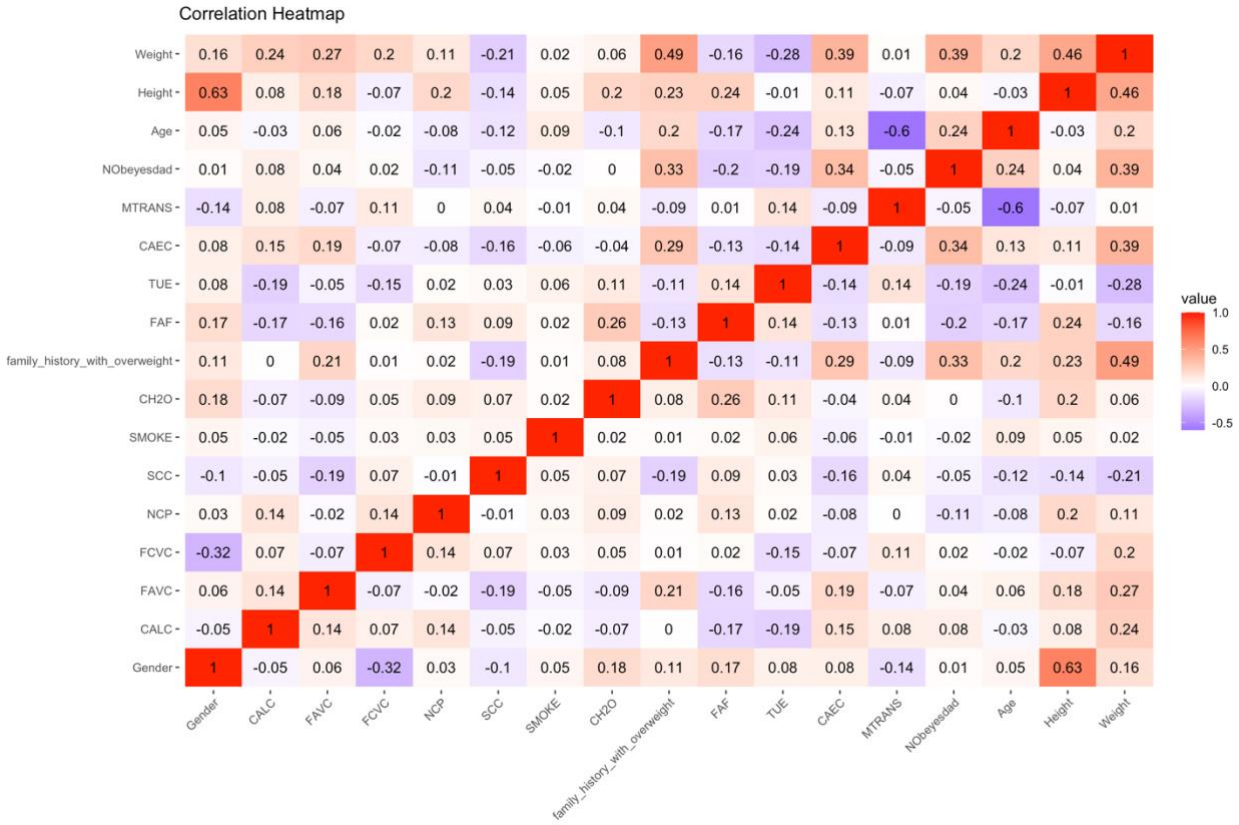


Correlation Matrix

A correlation matrix provides insight into the linear relationship between two features. Even though we will utilize a decision tree model for this project, understanding the correlations between features remains valuable is still important.

From the heatmap of the correlation matrix, we can observe that the obesity level exhibits a moderately strong positive correlation with several features. Firstly, it has a notable positive relationship with weight, indicating that as an individual's weight increases, their likelihood of being classified into a higher obesity level category also rises. Additionally, the obesity level shows a positive correlation with age, suggesting that older individuals tend to have a higher propensity for being in the overweight or obese categories.

Furthermore, the correlation matrix reveals a positive association between obesity level and the CAEC feature, which stands for "Do you eat any food between meals?". This implies that individuals who frequently consume snacks or meals between their main meals are more likely to



fall into higher obesity level categories. Lastly, the obesity level demonstrates a positive correlation with family history with overweight, suggesting that genetic or environmental factors related to being overweight within a family can contribute to an increased likelihood of obesity.

It's important to note that while these correlations provide insights into potential relationships, they do not necessarily imply causation. Further analysis and research would be required to establish causal links and understand the underlying mechanisms driving these associations. Nevertheless, the correlation matrix serves as a valuable tool for identifying potential risk factors and areas of focus in addressing obesity-related issues within the dataset.

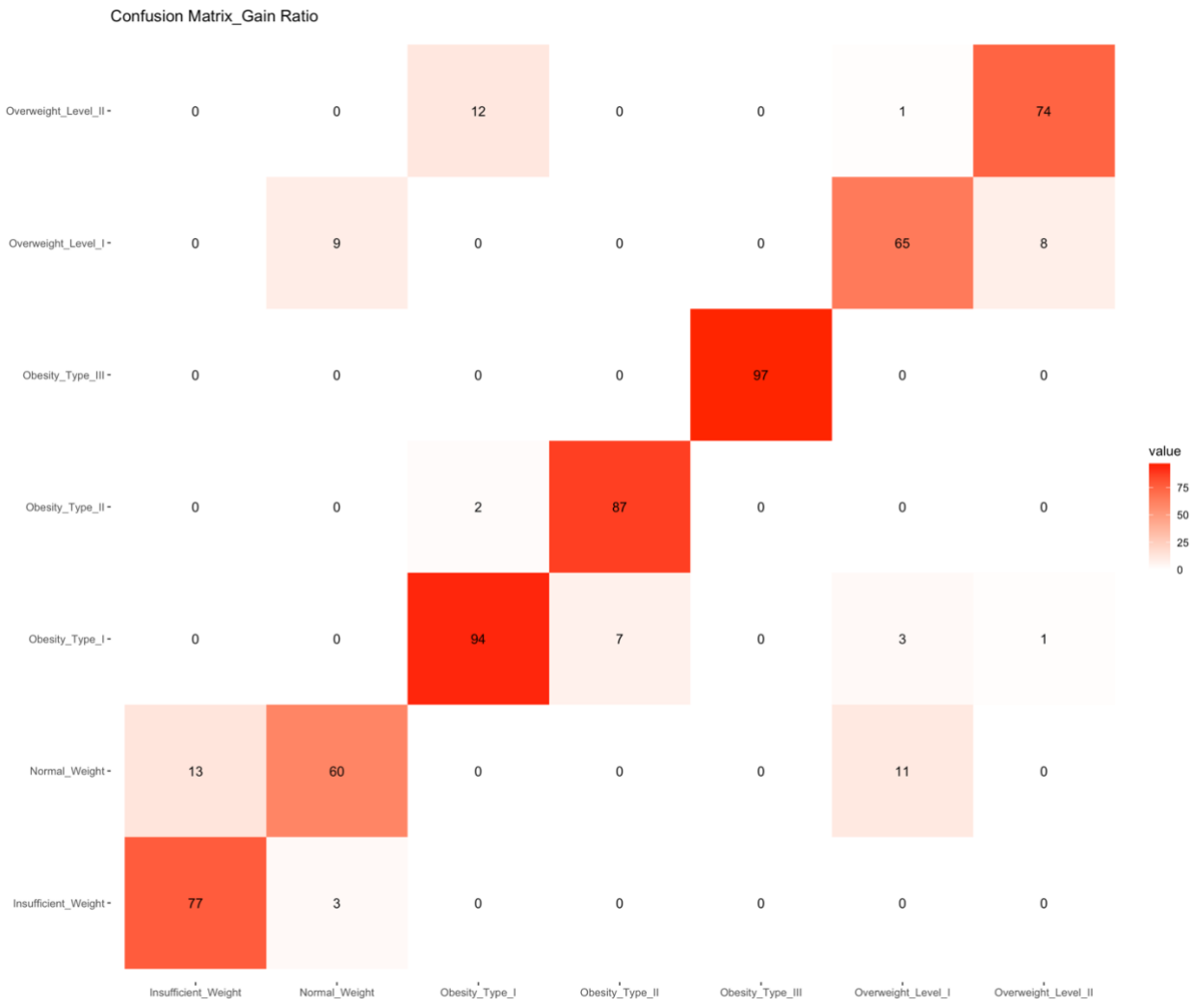
Decision Tree Models

In this project, we employed two different decision tree methods for constructing our predictive models: the gain ratio approach and the gini index approach. These methods are commonly used in decision tree algorithms to determine the best way to split the data at each node of the tree. The gain ratio method is an extension of the information gain method, which is based on the concept of entropy. While the information gain method solely considers the reduction in entropy or impurity, the gain ratio method also considers the intrinsic information or the number of subsets resulting from a split. It calculates the gain ratio by dividing the information gain by the intrinsic information, effectively normalizing the information gain by the split's intrinsic information.

On the other hand, the gini index method is a different approach for measuring the impurity of a set of examples. It calculates the Gini index for each possible split and selects the split that

results in the lowest Gini index value, indicating the highest purity or homogeneity of the resulting subsets. The Gini index ranges from 0 to 1, where 0 represents a completely homogeneous set (all examples belong to the same class), and 1 represents a completely heterogeneous set (examples are equally distributed among all classes).

Confusion Matrix

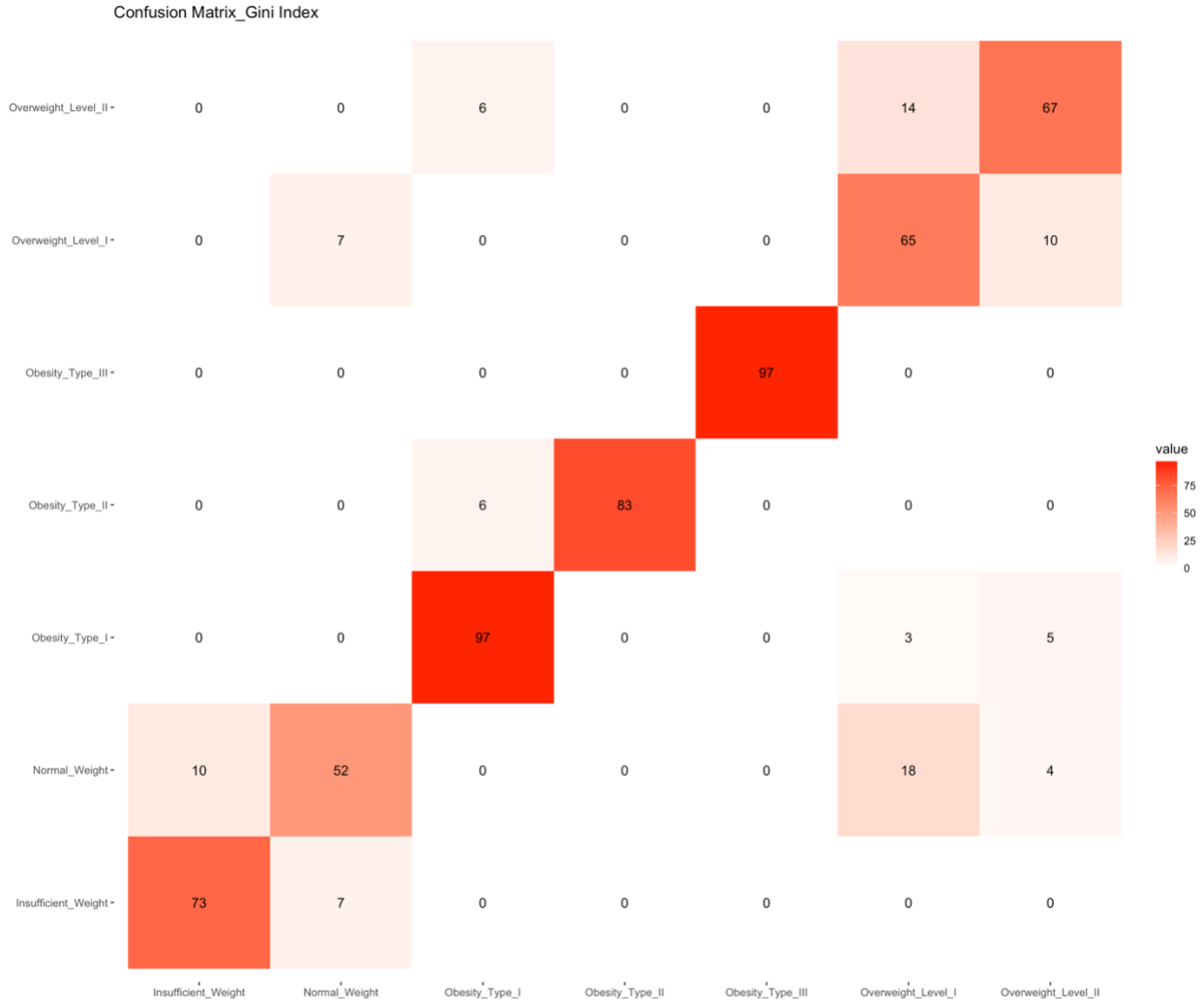


Evaluating the performance of classification models, especially for multi-class problems, can be challenging. In this project, where the target label has seven levels, the confusion matrix serves as a valuable metric for assessing the model's performance. The confusion matrix provides a

comprehensive view of how the model is performing by displaying the predicted target levels against the true target levels.

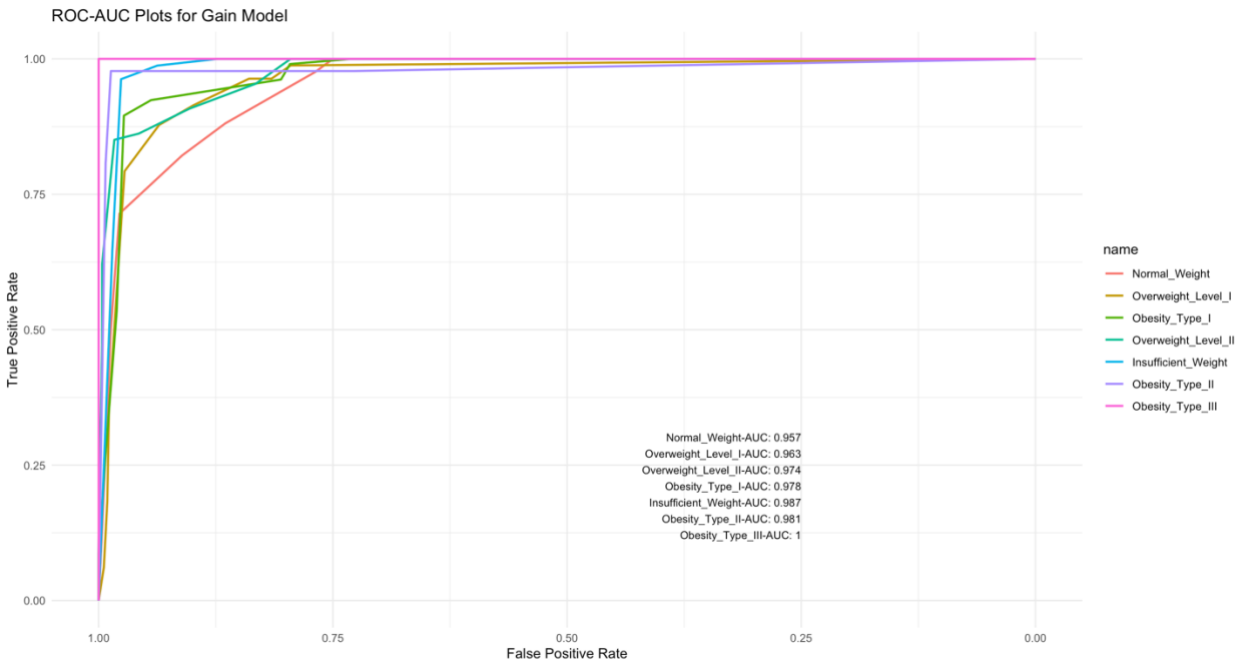
The x-axis of the confusion matrix represents the predicted target levels, while the y-axis represents the true target levels. Since we are dealing with a multi-class problem, the confusion matrix becomes more complex, as it needs to represent the predictions and ground truth for all seven levels simultaneously.

We have included two heatmaps of the confusion matrices, one for the model trained using the gain ratio method and another for the model trained using the Gini index method. Upon visual inspection, we do not observe significant differences between the two confusion matrices, suggesting that the overall performance of the two models is relatively similar.



However, quantitative metrics can provide a more precise comparison. The overall accuracy for the model trained using the gain ratio method is 0.8878, which means that approximately 88.78% of the instances were correctly classified across all seven target levels. On the other hand, the overall accuracy for the model trained using the Gini index method is slightly lower at 0.8558, indicating that around 85.58% of the instances were correctly classified.

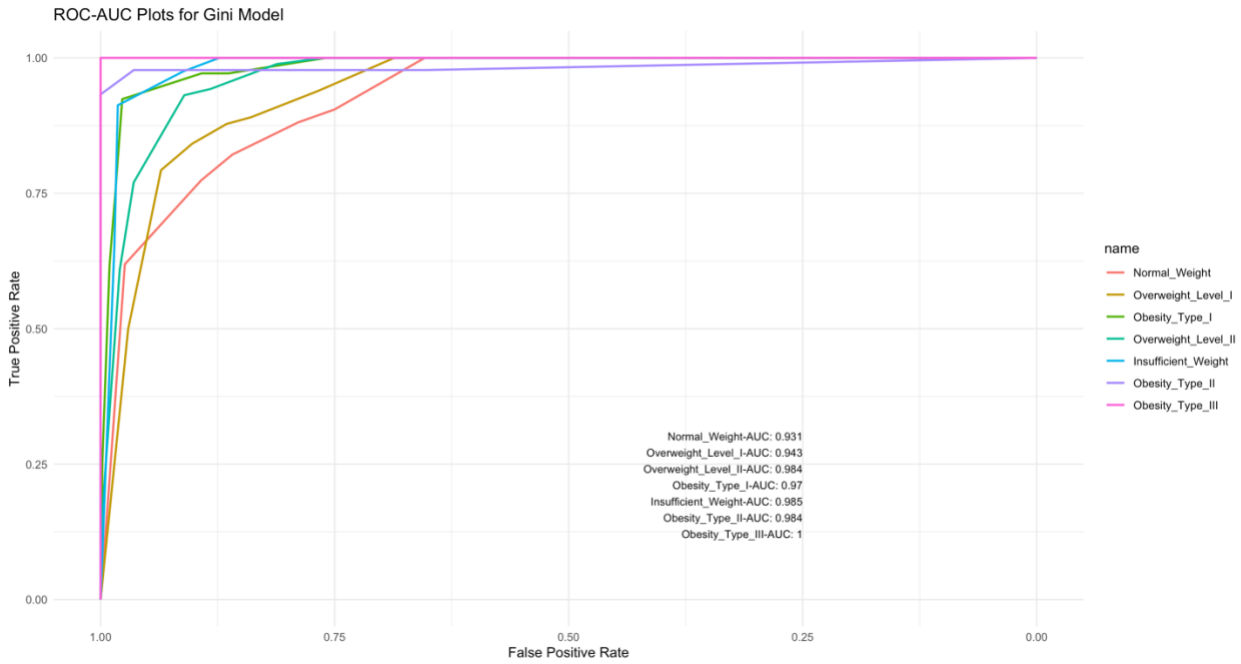
ROC-AUC Curve



In addition to the confusion matrix and overall accuracy, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is an important metric to evaluate the performance of multi-class classification models. The ROC-AUC provides a comprehensive assessment of the model's ability to distinguish between each class and the other classes combined.

Since our target label has seven levels, we would have seven individual ROC curves, each representing the performance of the classifier for a specific class compared to all the other classes combined. The ROC-AUC score ranges from 0 to 1, with a higher value indicating better

performance in discriminating between the positive class (the class of interest) and the negative class (all other classes combined).



Analyzing the ROC-AUC scores for both the gain ratio method and the Gini index method can provide valuable insights into their respective strengths and weaknesses in handling different classes within the multi-class problem.

For instance, if we observe that one method has a consistently higher ROC-AUC score across all classes compared to the other method, it suggests that the former method has superior overall performance in distinguishing between the classes. However, if the ROC-AUC scores vary across classes for the two methods, it may indicate that one method performs better for certain classes, while the other method excels at different classes.

Model Difference

While the overall performance metrics, such as accuracy and confusion matrices, suggest similar performance between the models trained using the gain ratio method and the Gini index method, there are notable differences in how these two approaches process and interpret the data.

One significant difference lies in the number of nodes in the resulting decision trees. The model trained using the gain ratio method has 191 nodes, which is substantially higher than the 95 nodes in the model trained using the Gini index method. This difference in tree complexity can have implications for the interpretability, generalization capabilities, and computational efficiency of the models.

A larger number of nodes typically indicates a more complex decision tree, which may capture intricate patterns and relationships in the data more effectively. However, this increased complexity can also lead to overfitting, particularly if the tree becomes too specific to the training data and fails to generalize well to new, unseen instances. On the other hand, a smaller number of nodes, as in the case of the Gini index model, may result in a simpler and more interpretable tree, but potentially at the cost of missing out on some nuanced patterns in the data.

Another notable difference between the two models lies in the variable importance scores assigned to the features. Variable importance measures the relative contribution of each feature to the model's predictions, providing insights into the most influential factors for the classification task.

Conclusion

In conclusion, this project compared two widely used decision tree methods - the gain ratio approach and the Gini index approach - for a multi-class classification problem with seven target levels related to obesity levels. While the overall performance metrics, such as accuracy and confusion matrices, showed relatively similar results between the two models, further analysis revealed notable differences in their inner workings and interpretations.

Despite using the same set of features, the gain ratio model and the Gini index model may assign different importance scores to the same features. This discrepancy arises from the distinct criteria used by each method to evaluate and split the data at each node of the decision tree.

The gain ratio method considers both the reduction in entropy (information gain) and the intrinsic information of the split, while the Gini index method focuses solely on minimizing the Gini impurity index. As a result, the two methods may prioritize different features or identify different combinations of features as being more informative for the classification task.