**Final Report: Predictive And Descriptive Analysis Results for**

**the Rio Grande Valley Obesity Study**

**(Final Term Project)**

**Submitted to:**

Prof. Francis Andoh-Baidoo

INFS 4330-01-Fall 2023

**Report Prepared By:**

Belen Rivas

Mac Franklin

Jorge Rodriguez

Juan Huerta

Odudu Otu

Ash Brooks

November 30, 2023

**Introducing Our Problem Definition:**

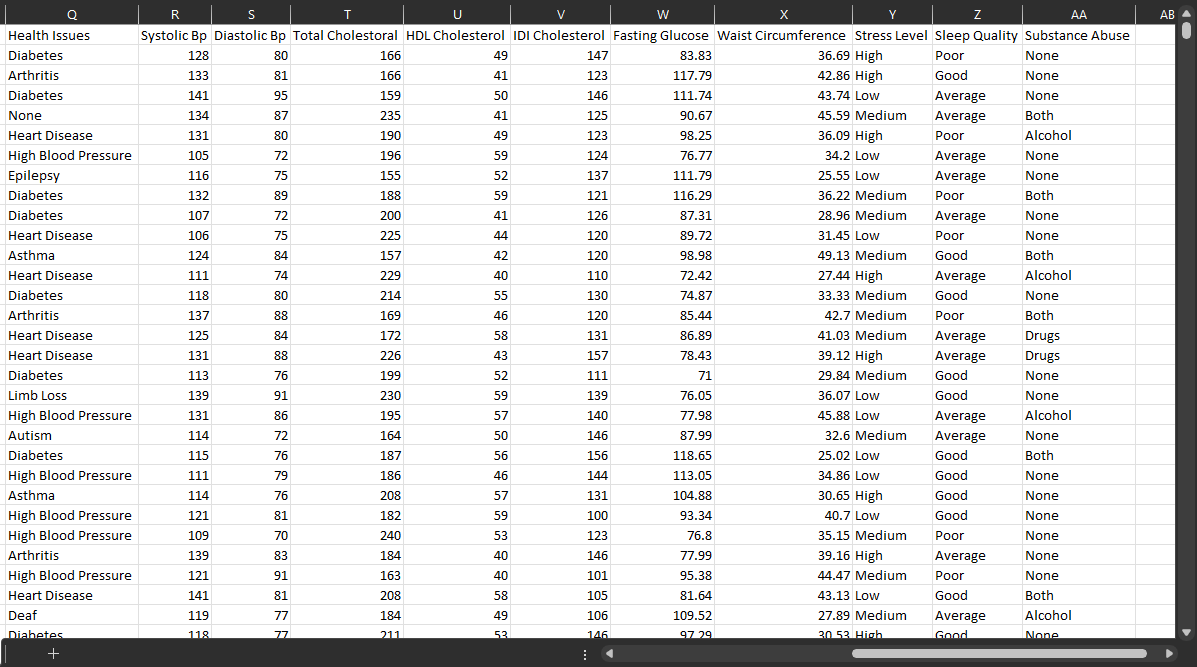
At the heart of the Rio Grande Valley, a public health challenge of considerable magnitude looms - the most alarming rates of obesity in America. As historically seen, this phenomenon is not a mere statistical anomaly, but an actual prolonged confluence of specific health, lifestyle, and socioeconomic factors that have been perilously shifting the fabric of the Rio Grande Valley community. Our final project is a determined foray into unraveling the acute reality of obesity in this region through the utilization of various data mining and machine learning techniques that will allow us different - yet as well interesting - perspectives of the critical situation. At the very core of our study lies our developed and targeted question: What factors contribute to the high presence of obesity, and how can we determine whether someone is obese or not? This question further branches off into specific and selected analytical assessments, such as assessing the role of demographics in obesity rates, understanding the relationship between Body Mass Index (BMI) and obesity, delving into the behavioral patterns that correlate with obesity, and assessing how ethnicity and economic status intersect with obesity prevalence. Lastly, we are as well readily focusing on whether we can segment the population at risk of being obese. From framing our investigation through these certain objectives and inquiries, we will ensure a deeper exploration into the obesity crisis that shadows the communities of the RGV.

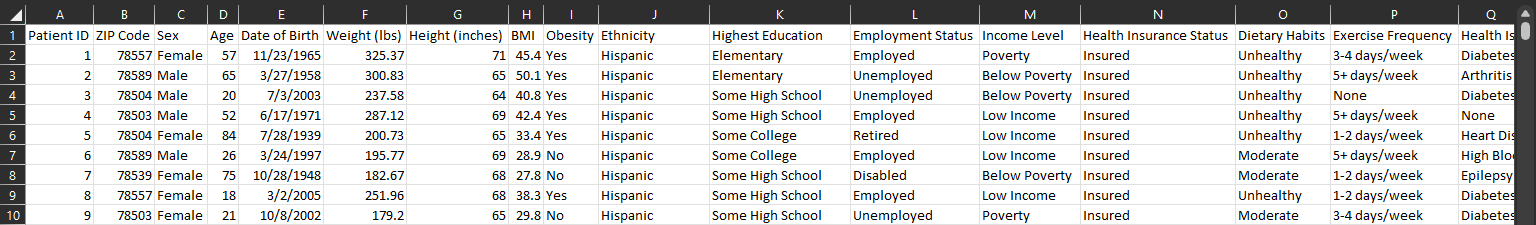
**Showcasing a Business Understanding:**

As we know that a business understanding of our problem is pivotal in successfully furthering or progressing within our research, it’s crucial to note that our study aligns with critical policy-making and patient care. For instance, from the various insights that we will gather from our models and algorithms, we could reinforce certain knowledgeable information and findings that policy makers or public health officials could use when formulating and implementing healthcare policies or initiatives. Health providers could as well use our fortifying insights when assessing what preventative strategies and treatment plans would most resonate with the needs of the Rio Grande Valley community and its people. This could all substantially impact their overall quality of life and societal well-being. Additionally, with a stronger focus on obesity and a heightened mission of diminishing its prevalence, an immense decrease of healthcare costs could ensuingly occur. Thus, addressing obesity is both a socially and economically significant pursuit that entails far-reaching implications for a range of stakeholders within the RGV.

**Showcasing Data Understanding:**

Our comprehensive dataset, which draws information from 1,000 samples/individuals, offers a broad and almost encyclopedic view of the variables potentially influencing obesity in the RGV. For instance, it includes necessary demographic information, such as ethnicity, age, gender, and countless more, which are all crucial in mapping the distribution of obesity across different population segments. Health metrics such as weight, height, BMI, and obesity status are found in the dataset and form the very core of our health-related analysis, since they provide the insights into the physical health indicators that directly help in solving our leading data analytics question. Lifestyle indicators are as well included and vital to our study, as factors such as dietary habits, exercise frequency, and stress levels, shed light on the important daily behaviors and choices of individuals. Exploring these patterns could help our study imperatively highlight the impact of different lifestyles on obesity and make a connection between the ingrained cultural lifestyles of the RGV communities and the rising health issue within the associated regions. The dataset as well explores various pivotal socioeconomic dimensions, such as income level, highest education, employment status, and more. Lastly, the dataset additionally includes numerous clinical and psychological variables, which invaluably provide both an in-depth physical and mental health profile for each individual. At large, each and every column found in our file is consequential towards our study, as each demonstrates an evidently strong link with the obesity status variable. This variable, located within column i, is our outcome variable for our predictive modeling endeavors. It is the most decisive and key element within our conducted

Research.



**Explanations of Chosen Techniques**

*(1) How do obesity rates differ across various demographic groups such as ethnicity, and socioeconomic status?*

Methodology: Visualization Utilizing Excel & Tableau

Rationale: Visualization is a profound tool for deriving and displaying certain patterns and insights with data, especially when dealing with diverse demographic groups. By utilizing Excel and Tableau, we can directly create and reveal visual representations that informatively highlight variations in obesity rates across different ethnicities and socioeconomic statuses. These visualizations and the graphical nature of their approach enables us to all-inclusively and easily identify disparities within our dataset that effectively exhibits the distribution and variation of obesity rates across our desired factors. It also enables us to isolate certain patterns that we wish to make visually accessible, which is suitable for the question at hand.

*(2) Is BMI directly related to obesity? What are the factors that predict bmi?*

Methodology: Decision Tree

Rationale: We chose the predictive analytics model of a decision tree for this question, as we believed it was an ideal choice for examining the relationship between BMI and obesity due to its ability to model hierarchical decision-making processes. Using SPSS for this supervised learning approach allows us to identify the most significant predictors of BMI and how these factors interact to predict obesity. With decision trees being interpretable through decision-making paths, if-then relationships between variables, and resulting outputs of predictive accuracy, we can more easily and conclusively understand the direct and indirect influences on BMI and, subsequently, obesity. This also would enable us to straightforwardly determine what metrics would make an individual obese or not through the resulting predictors of BMI, furthering our investigation into answering the central inquiry within this study.

*(3) Are there any behavioral patterns that correlate with the obesity rates in our dataset?*

Methodologies: K’s Nearest Neighbor, Random Forest Tree

Rationale:

In this study, we chose the K-Nearest Neighbor (KNN) and Random Forest Tree methodologies for their distinct analytical strengths, tailored to our research goals. KNN is adept at identifying subtle behavioral patterns, particularly useful in discerning lifestyle elements like stress, exercise, and diet among individuals with similar obesity profiles. Complementing KNN, we incorporated Random Forest Tree, a technique known for its capability to assess various behavioral characteristics collectively. This method provides a holistic view of the factors contributing to obesity, enhancing the accuracy and depth of our study. By integrating these techniques, we aim to conduct a comprehensive examination of the behavioral patterns in our dataset and their correlation with obesity rates. This combination allows for a more detailed and insightful exploration of the complex dynamics underlying obesity.

*(4) If someone is Hispanic and has a low poverty level, how often will they be obese?*

Methodology: Association Rule Mining

Rationale: In addressing this question, we selectively favored the technique of unsupervised association rule mining. Its procedure of identifying sets of elements that recurrently appear together in datasets and generating association rules from them is the procedure we wished to undertake for the characteristics of (1) being Hispanic, (2) being low income and (3) being obese. Deriving association rules based on these sets of characteristics would provide us with the necessary insights on the relationship between ethnicity, poverty level, and obesity rates. This is overall crucial in determining how certain socioeconomic and ethnic factors are associated with obesity.

*(5) Can we segment the population at risk of being obese?*

Methodology: Cluster Analysis

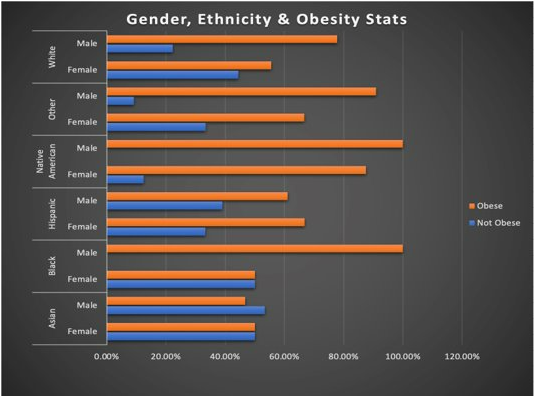
Rationale: In response to this question, we chose the unsupervised learning technique of cluster analysis, as this method enables us to categorize the population into distinct groups based on their shared health and lifestyle attributes. By applying cluster analysis to our dataset, we could identify patterns and commonalities in relevant health factors. It would help us isolate specific segments of the population that exhibit characteristics or behaviors that put them at a higher risk for obesity. Furthermore, the specific ability to segment the population in this manner is crucial for developing more targeted approaches to obesity prevention and management. It allows for the creation of intervention strategies that are specifically established to address the unique needs and risk factors of each identified cluster, thereby enhancing the effectiveness of public health strategies aimed at combating obesity. Thus, this tailored approach is instrumental in addressing the varied nature of obesity risk factors within the diverse population of the Rio Grande Valley.

**Interpretation of Results from Chosen Techniques**

*(1) How do obesity rates differ across various demographic groups such as ethnicity, and socioeconomic status?*

Software Used: Excel & Tableau

Output:





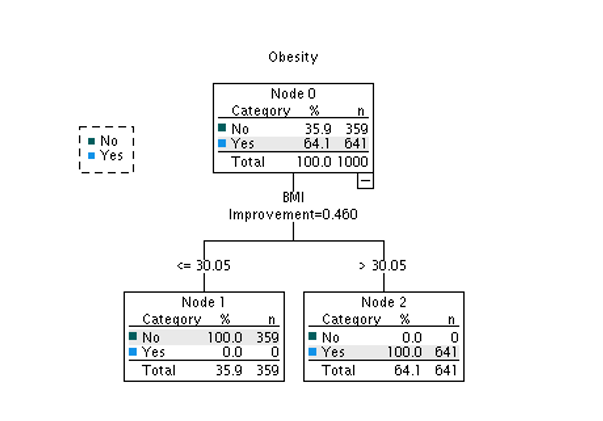
Interpretation:

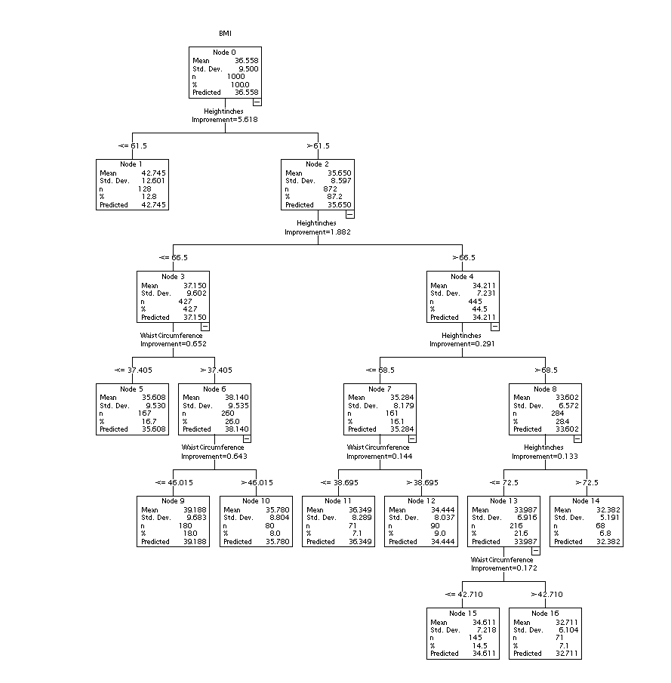
The first bar graph directly responds to the portion of the question of how obesity rates differ among various ethnic groups by showing a clear comparison across demographic populations. Specifically, it highlights that within each ethnic group, there are distinct levels of obesity prevalence, with the Hispanic population showing a particularly high rate in comparison to some of the other ethnicities presented. This reinforces that ethnicity is a significant factor in obesity prevalence and that the Hispanic group does require critical focus. Gender differences are noted within each ethnic category, but the key takeaway is the prominent variation in obesity rates between the different ethnic groups. It reflects the central concern of this question or Q1. Furthermore, this visual image is crucial for understanding which demographics might benefit most from developed obesity plans.

The second visualization dashboard displays variations across economic statuses such as education level and income level against obesity. To highlight the Hispanic population within these economic groups, we as well incorporated ethnicity within our charts. According to our charts, Hispanics dominate obesity rates, with those only having some high school education showing the most pronounced level of obesity. The Hispanic population in poverty also rank the highest within our chart of income level, ethnicity, and obesity.

*(2) Is BMI directly related to obesity? What are the factors that predict bmi?*

Software Used: SPSS

Output: 



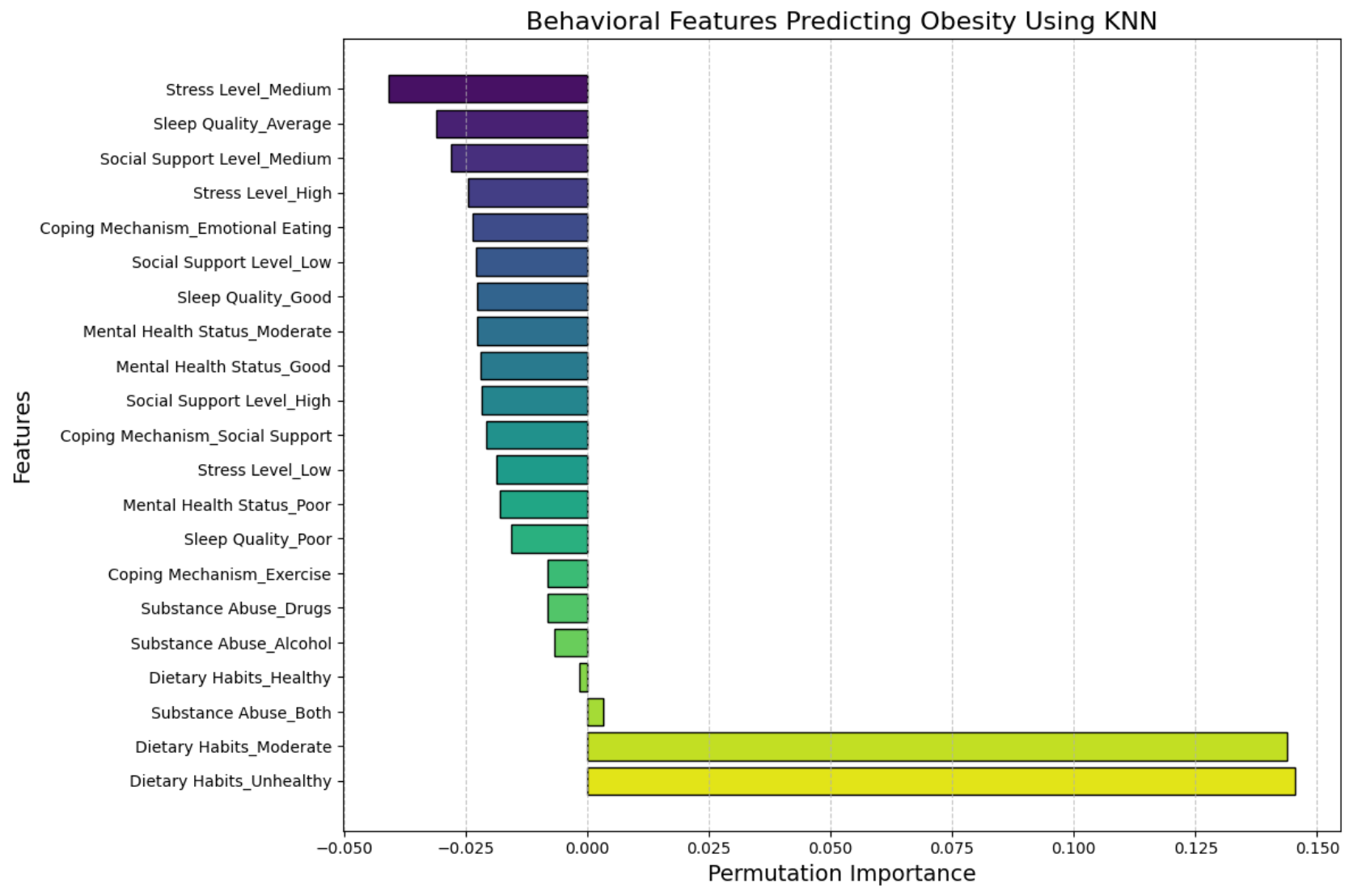
Interpretation:

The decision trees drawn and demonstrated above are instrumental in deciphering the relationship between BMI and obesity, as well in identifying the factors that directly predict BMI. The branches shown directly indicate that height and waist circumference are two of the most relevant predictors for BMI and are precisely critical in calculating an individual's body mass index. This reinforces the comprehension that BMI is derived from these two measurements. Furthermore, the nodes within the tree show a clear decision-making pathway where specific ranges of BMI are used to predict obesity, affirming the direct relationship between BMI and obesity status. The classification tree simplifies this further by dichotomizing the outcome into obese (Yes) and not obese (No) based on a BMI threshold of 30.05, a standard cut-off within the medical community. The tree confirms that individuals with a BMI above this threshold are consistently categorized as obese, validating BMI as a strong predictor of obesity. In answering the leading question of this study, this data strongly supports the conclusion that a higher BMI is indicative of obesity, providing us with a reliable metric for determining whether someone is obese or not. These findings as well stress the importance of monitoring BMI as part of health assessments in the RGV and suggest that supporting a healthy height-to-waist ratio could be effective in addressing obesity.

*(3) Are there any behavioral patterns that correlate with the obesity rates in our dataset?*

Software Used: Python

Output:



This graph provides a visual representation of the permutation feature importance of various behavioral features in predicting obesity based on a K-Nearest Neighbors (KNN) model.

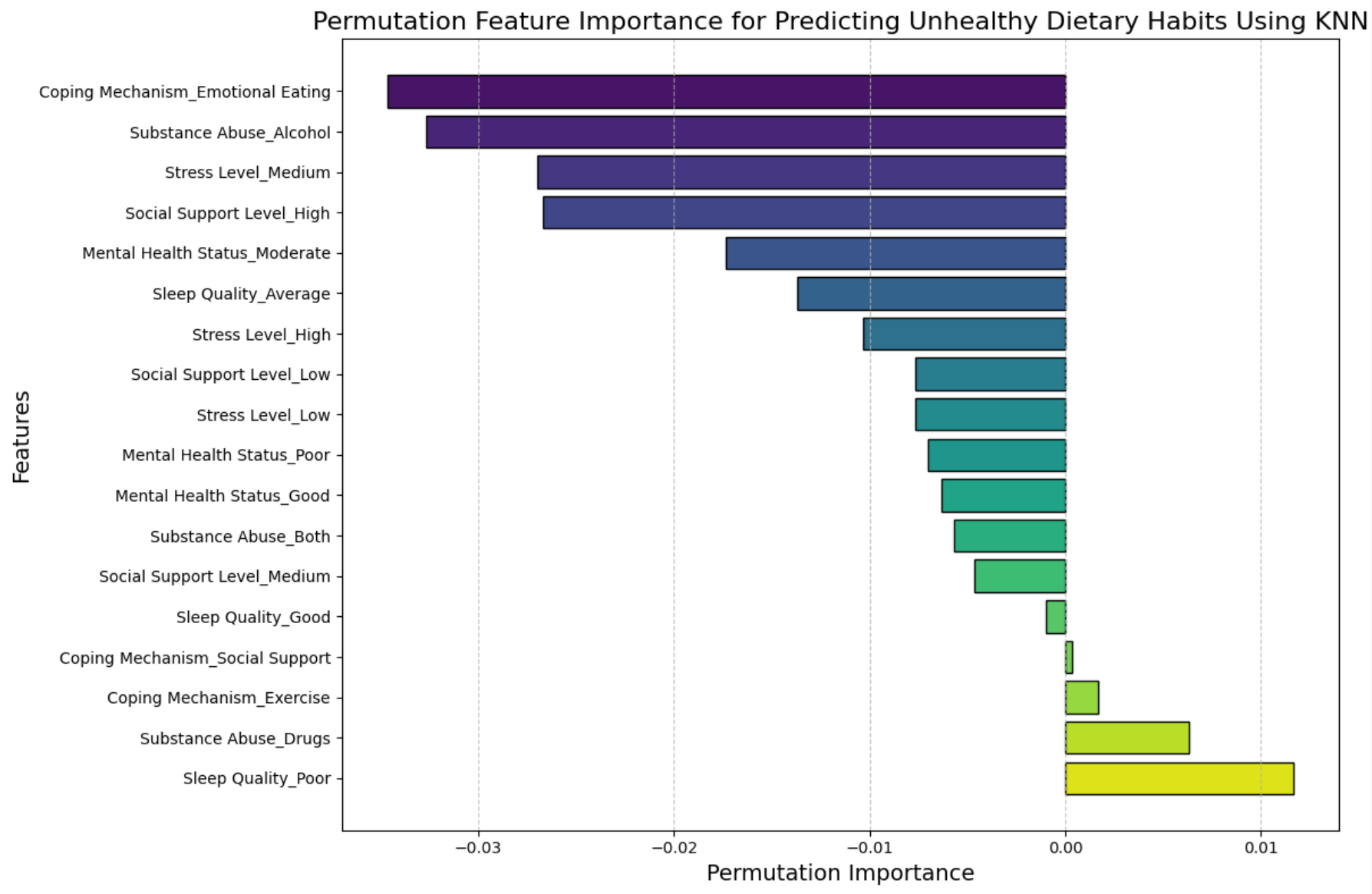
Graph Observations:

Dietary Habits**:** Among the features, dietary habits emerged as a significant factor. Particularly, 'Dietary Habits\_Unhealthy' was identified as the most influential predictor of obesity.

Substance Abuse**:** Our model indicates that substance abuse, including drug and alcohol use, moderately contributes to higher rates of obesity.

Mental Health and Coping Mechanisms**:** The data reveals a spectrum of mental health states and coping strategies, with varying degrees of impact. Notably, 'Coping Mechanism\_Emotional Eating' shows a significant relationship with obesity, suggesting a substantial connection between stress-coping methods and obesity.

Social Support and Stress Levels: We observed a correlation between obesity and social dynamics. Lower social support and higher stress levels are indicative of a higher risk of obesity.



The following graph displays the behavioral permutation feature importance for predicting unhealthy dietary habits using (KNN), showcasing our interest in seeing what determinants influence that highly significant factor with substantial predictive power on obesity.

Graphing Observations:

Coping Mechanisms: Coping mechanisms, especially 'Coping Mechanism\_Emotional Eating', stand out as the most significant factor. This suggests that how individuals cope with stress or emotional distress may have a considerable impact on their dietary habits.

Substance Abuse**:** Substance abuse, particularly alcohol, is shown to have a notable importance. This indicates a potential link between alcohol consumption and unhealthy dietary habits.

Stress Levels**:** Different levels of stress ('Stress Level\_Medium', 'Stress Level\_High', 'Stress Level\_Low') have varying degrees of importance. The model suggests that stress levels could influence dietary choices, with 'Stress Level\_Medium' having a higher importance.

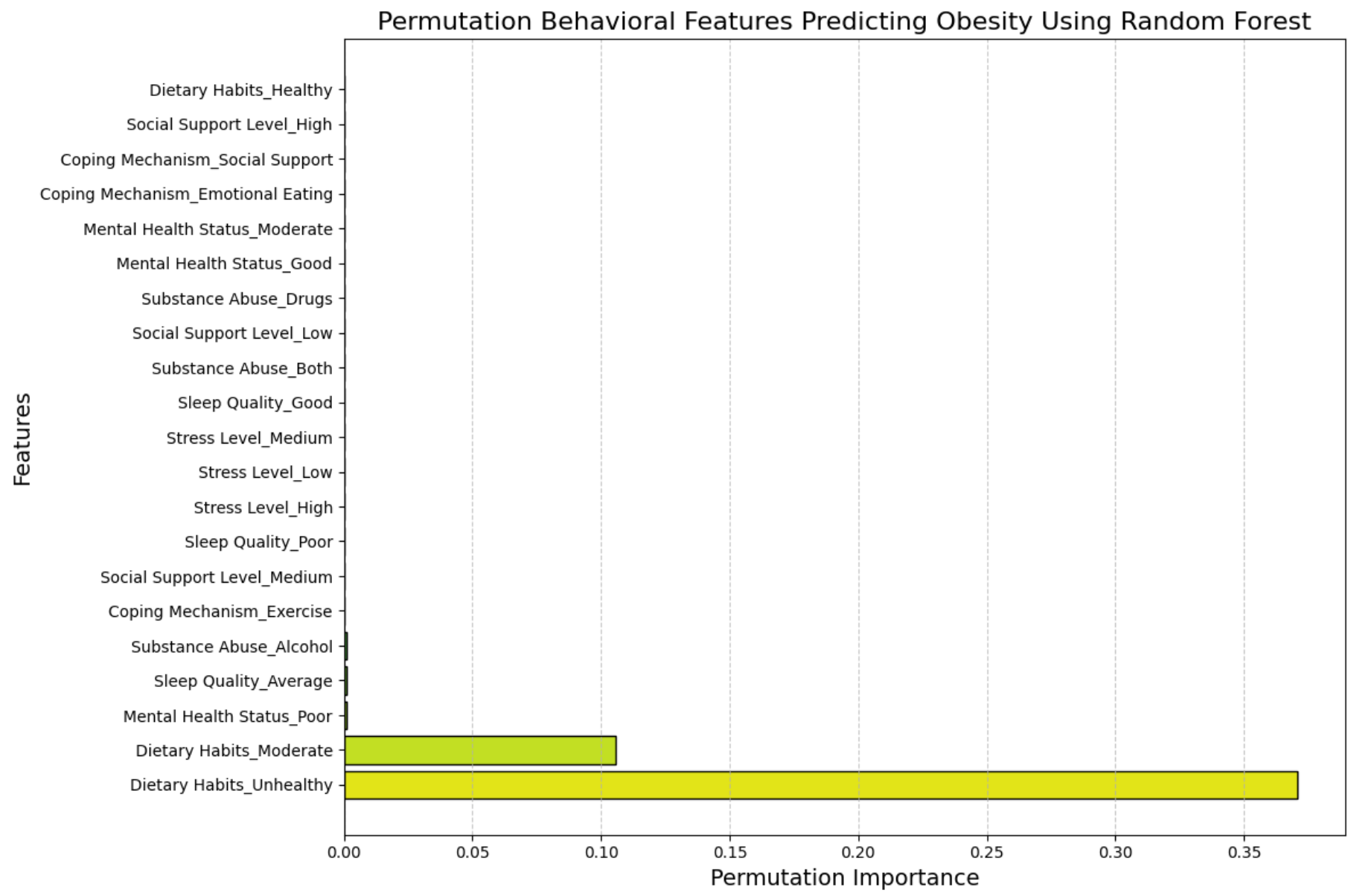
Social Support**:** The levels of social support, from 'Social Support Level\_High' to 'Social Support Level\_Low', are significant predictors. This may imply that the presence or absence of a support system can influence eating behaviors.

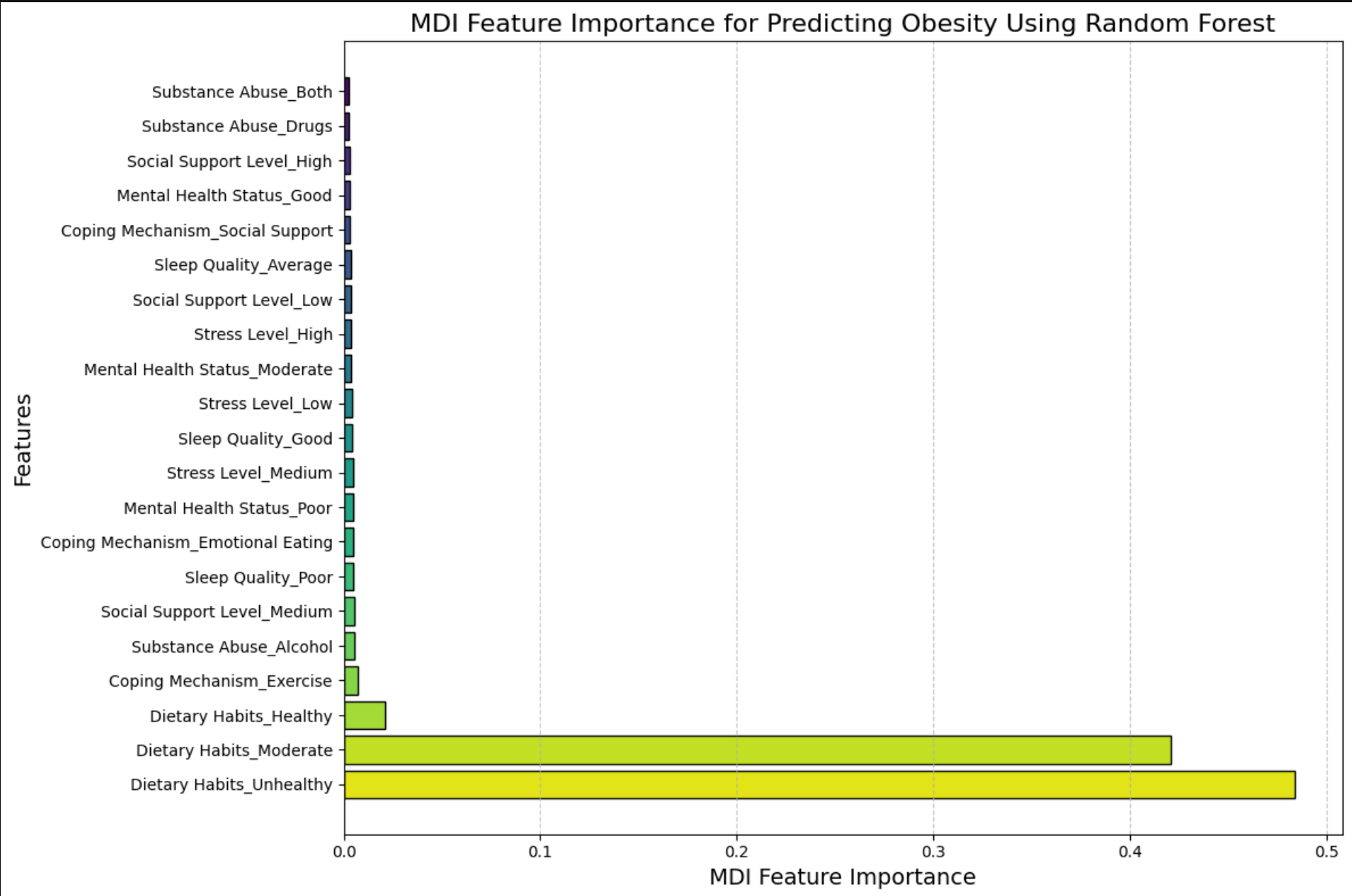
Mental Health Status**:** The mental health status of individuals, particularly 'Mental Health Status\_Good', appears to be a predictive factor.

Sleep Quality**:** Sleep quality, both 'Sleep Quality\_Average' and 'Sleep Quality\_Good', is identified as important, which aligns with research suggesting that sleep patterns can affect eating habits.

Negative Values**:** Some features have negative importance scores, suggesting that they’re not as useful for the prediction or are redundant given the other features in the model.

Through Random Forest Trees, we aim to delve deeper into the significance of each behavioral feature, potentially unveiling patterns and interactions that the KNN model might have missed





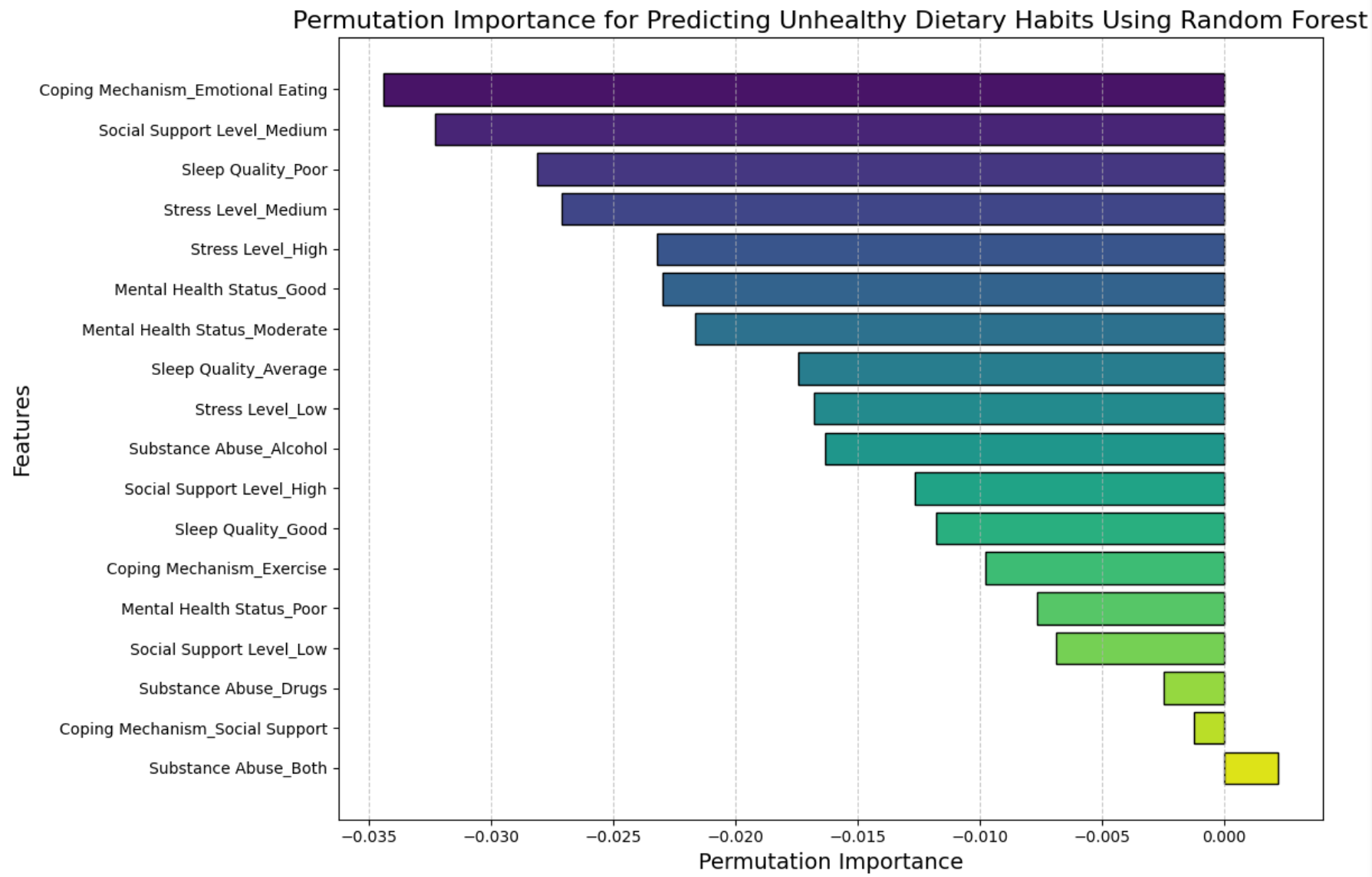
In this analysis, we applied permutation importance, as previously done with KNN, and also employed Mean Decrease in Impurity (MDI), a technique specific to Random Forest Trees. MDI quantifies a feature's importance based on its contribution to the purity of decision nodes within the trees of the forest.

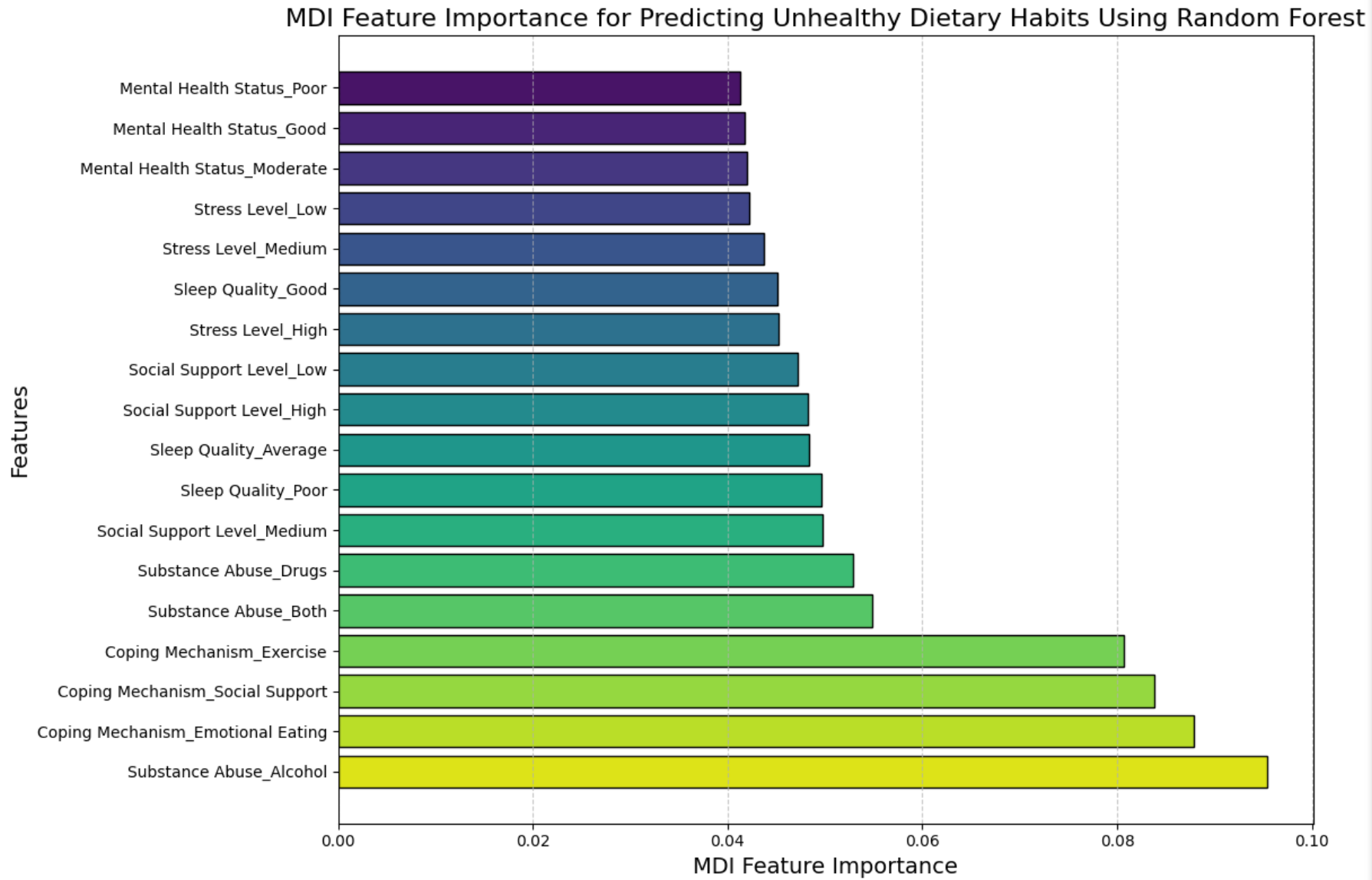
Graph Observations:

Consistency Across Both Methods**:** 'Dietary Habits\_Unhealthy' appears as the most important feature in both analyses, indicating a strong and consistent relationship between unhealthy dietary habits and obesity.

Variability in Importance**:** Other features show variability in their importance between the two methods. For example, 'Dietary Habits\_Healthy' is more important in the permutation importance analysis compared to the MDI analysis. This could suggest that while healthy dietary habits are predictive, they may not always be the best at improving model purity.

Influence of Lifestyle Factors: Lifestyle factors like 'Substance Abuse\_Alcohol', 'Coping Mechanism\_Emotional Eating', and 'Sleep Quality' vary in their ranked importance between the two methods.

Impact of Stress and Social Support**:** Stress levels and social support appear to have a significant impact on obesity prediction, with their rankings varying between the two methods. This could indicate that these features are intertwined with other features, influencing their permutation importance.



The two graphs display the feature importance of various behavioral factors in predicting unhealthy dietary habits, as we did with our KNN models. Now we use the Random Forest Tree model in discerning what influences the dietary habits variable. One graph illustrates permutation importance, while the other shows (MDI) feature importance.

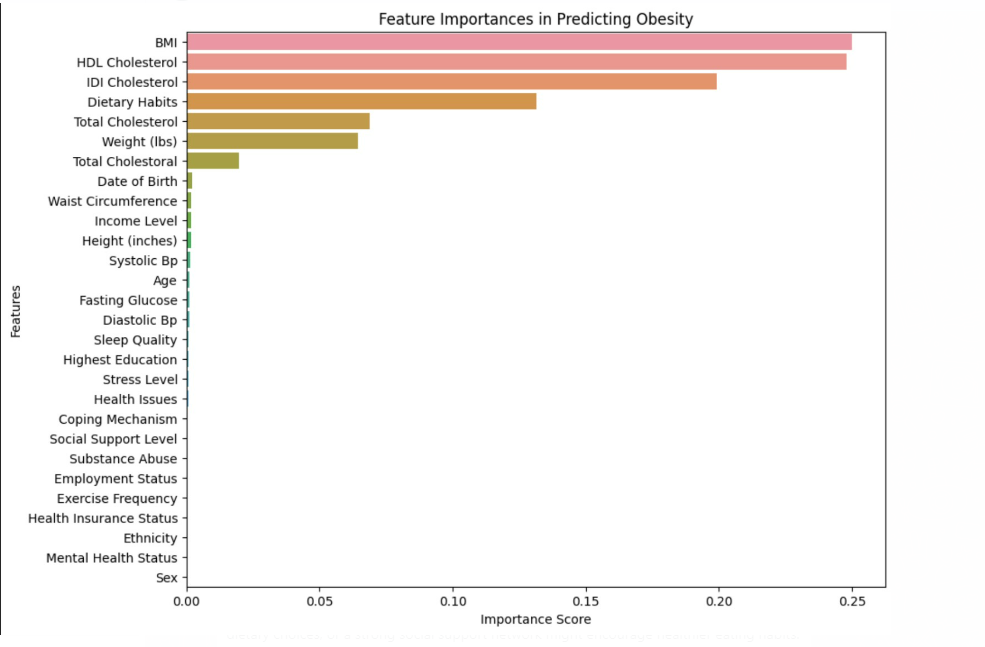
Graph Observations:

Predictive Value**:** MDI suggests that emotional eating, alcohol abuse, and poor sleep quality are particularly relevant for predicting unhealthy dietary habits, potentially due to their strong individual contributions to the model's decisions.

Differing Importance Scores**:** The discrepancy between the importance scores in permutation importance and MDI could be attributed to the methods' differing sensitivities to feature interactions and the dataset's inherent noise.

Model Interpretation: The contrasting results between permutation and MDI importance underscore the importance of using multiple methods for feature importance to get a more comprehensive understanding of the factors at play.

Further Analysis**:** Further investigations could explore the relationships between these features and unhealthy dietary habits, considering the possibility of interactions between them that could affect their predictive power.



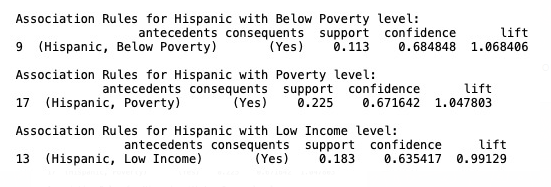
Interpretation:

The provided output and graphs give a thorough understanding of the association between various behaviors and obesity. They particularly highlight the significant role of dietary habits, which exhibit the highest correlation with obesity among all behavioral factors. This underscores the direct impact of diet on obesity. Additionally, sleep quality, stress levels, and substance abuse are also highly correlated with obesity. These findings suggest that modifying behaviors, such as improving dietary and sleep habits or reducing substance use, could significantly influence obesity outcomes. Furthermore, the feature importance charts from KNN and RFT models deepen our understanding by ranking a wide array of obesity predictors, again emphasizing dietary habits as key contributors. The charts also illustrate that behavioral patterns are crucial in mitigating obesity risk, with factors like unhealthy eating and alcohol abuse being noteworthy. Moreover, the last chart's focus on BMI addresses our core question about determining obesity. BMI, as a combined measure of weight, width, and height, is a standard obesity criterion. Its predictive strength in the chart confirms its reliability as an obesity indicator. Overall, these results confirm the significance of various behavioral patterns in relation to obesity and establish BMI as the most critical determinant of obesity, the study's target variable.

*(4) If someone is Hispanic and has a low poverty level, how often will they be obese?*

Software Used: Python

Output:



Interpretation:

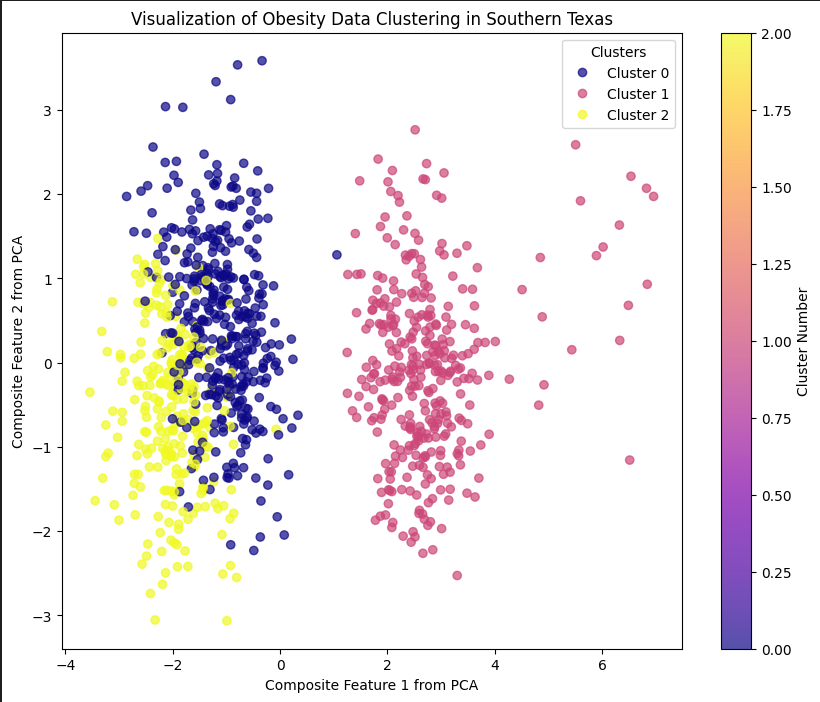
The output from our association rule mining uncovers informative insights regarding the existing relationship between an individual's income level, ethnicity, and obesity status. The rules generated uncover substantial associations between being Hispanic with either a low income level, a general poverty level, and a below poverty level and the likelihood of being classified as obese. The confidence level of .635417 suggests that, among the Hispanic population with a low income level in our data, there is a 63.54% probability of obesity. For Hispanics with a general poverty level, there is a 67.16% probability of obesity. Lastly, for Hispanics with below poverty level, the confidence level of 0.684848 suggests that there is approximately a 68.48% chance of obesity.These levels of confidence signifies strong correlations between the antecedent conditions and the consequent classification of obesity.

This reveals that, when a Hispanic individual within our study is also experiencing low income, they will be obese more than half the time.

*(5) Can we segment the population at risk of being obese?*

Software Used: Python

Output:



Interpretation:

The scatter plot above illustrates a clear segmentation of the population into clusters that reflect varying levels of obesity risk, effectively answering the question at hand. The three clusters differentiated by color are: (1) Cluster 0 (Blue): This cluster suggests a demographic with moderate to high BMI, which indicates elevated risk of obesity. The dense grouping within this cluster points to shared characteristics that may influence obesity risk, such as similar dietary habits, exercise frequencies, or other health indicators. (2) Cluster 1 (Pink): Individuals in this cluster tend to have lower BMI values, signifying a lower risk of obesity. The spread of the data points could imply a wider variation in lifestyle behaviors or health attributes that correlate with a healthier weight range. (3) Cluster 2 (Yellow): This cluster is distinguished by the highest BMI values, signaling a segment of the population at the highest risk of obesity. The concentration of these points highlights the urgency for targeted health interventions for this group. Moreover, by identifying these clusters, we have effectively segmented the population according to risk factors associated with obesity. This segmentation is instrumental in enabling healthcare providers and policymakers to tailor their strategies and resources more effectively. For instance, Cluster 2, which exhibits the highest risk, might be prioritized for intervention programs aimed at weight management and lifestyle modification. Meanwhile, preventive measures and education could be the focus for Cluster 1, which represents a lower-risk group.

**Covering Supervised Learning:**

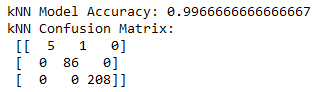
We applied supervised learning techniques like decision classification trees, K’s Nearest Neighbor Models, and Random Forest Tree programs. We tested and quantified the influence of behavioral patterns on our target variable of obesity status numerous times using various algorithms. The random forest trees and KNN in python specifically provided us with an extremely detailed and supplemented understanding of both behavioral patterns and other numerous variables that contributed to obesity. We confirmed and revealed the predictive power of all the behavioral factors in our dataset and further used our models to discover the influences on those highly significant factors themselves. We got to explore a plethora of interesting and meaningful relationships. Furthermore, the decision tree analysis, grounded in SPSS, was particularly revelatory, illustrating the ordered intricacies among various predictors. It helped us establish within our study the use of body mass index as the highest-ranking determinant of obesity itself. With high-levels of accuracy, we thus crafted visualizations that directly aligned with both our second and central data analytics question. This narrative not only informs the current state of obesity within the population but also equips us with the predictive power to identify individuals at risk. We have, with the use of supervised learning, advanced our ability to discern patterns and predict obesity.

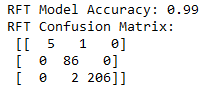
**Covering Unsupervised Learning:**

We undertook unsupervised learning through our exploration of patterns and relationships that are not explicitly labeled within our data but are strongly linked to obesity. In fact, through the lens of association rule mining and cluster analysis, we were able to discover associations and create population segments in ways we could not pursue within the umbrella of supervised learning. Association rule mining revealed how certain characteristics frequently converge among individuals with obesity. This technique drew out specific rules that indicated the likelihood of obesity given the presence of certain factors. Thus, we could isolate certain individual variables and determine an individual's probability of having obesity. Cluster analysis then furthered our research and study by grouping individuals into clusters based on shared attributes, effectively identifying subpopulations with varying degrees of obesity risk. By integrating these unsupervised learning techniques, we gained a much more elaborate view of the obesity epidemic.

**Accuracy and Effectiveness of Models:**

The effectively developed quantitative analysis we conducted through our models presents compelling evidence regarding the obesity crisis in the Rio Grande Valley. We can see the effectiveness and accuracy of these models if we turn to, for example, our K-Nearest Neighbors and Random Forest Tree algorithms. After some conducted testing, these models exhibited high accuracy scores with KNN at approximately 99.7% and RFT at approximately 99%. These scores, alongside the confusion matrices, indicate a strong predictive performance, with a significant number of true positive predictions and minimal false negatives or false positives. The high true positive rates suggest that our models are effectively capturing the complex interaction of factors that signal obesity presence. They provide tangible evidence and a theoretical reinforcement of the health crisis.





**Summary of Key Findings:**

Our data analysis has substantiated the critical factors contributing to obesity in the Rio Grande Valley. BMI has been found to be the primary indicator of obesity, with decision tree models highlighting its direct relationship to height and waist circumference. Behavioral influences, particularly dietary habits, have been presented and corroborated as strong determinants of obesity. Socioeconomic factors like income level, education level, and ethnicity, notably within the Hispanic community, have been identified as significant correlates to obesity through association rule mining and other visualization techniques. Cluster analysis further segmented the population, pinpointing specific risk groups. Overall, these gathered findings provided a substantial data-driven picture for our study of obesity.

**Final Outcome Interpretation - Solving the Core Problem:**

As we conclusively come to a final interpretation of our extensive data analytics project on obesity in the Rio Grande Valley, it becomes increasingly clear that the most prominent and pivotal question driving our research - What factors contribute to the high presence of obesity, and how can we determine whether someone is obese or not? - finds its answers in the large-scaled and varied data we analyzed. Through this analysis, we detected factors that largely contribute to the diagnosis of obesity for an individual: Dietary habits, poor sleep quality, alcohol or general substance abuse, high stress levels, poor coping mechanisms, poor mental health, poor social support, ethnicity, low education level, and low income level - all, to a great and concerning extent - with some more impactful than others, move an individual closer to obesity and, consequently, give rise to the high prevalence of obesity in the RGV. These factors constitute the broader pattern impacting the alarming community health of our targeted demographic. In terms of identifying obesity, our study confirms that BMI is a critical and almost direct indicator. An unhealthy height to waist circumference ratio will directly reveal an individual's obesity status - determining whether an individual is obese or not. This is consistent across three of our highly reliable models - decision tree, knn, and random forest tree, indicating a high level of reliability in BMI as a measure for determining obesity. As we look back at the business implications and aspects of our research, this knowledge and awareness could provide stakeholders, primarily health providers and policy-makers, with the reinforced information necessary to prompt new health strategies and patterns in the RGV. Such strategies could include added sources of stress management, added sources of social support, enhanced monitoring of substance usage, enhanced monitoring of BMI, more dietary food options, and more. It’s worth exploring the diversity of initiatives the RGV community could pursue to alleviate the findings that were drawn from this final term project.

**Code Snippets and Technical Details**

For technical details, we used a collaborative effort of technology and software applications including but not limited to Jupyter Notebook, Python, SPSS, excel, and more. Notebook was used a lot in correlation with Python to set up graphs, whereas excel handled data tables and SPSS gave us our decision trees. There will be code snippets down below.

We did NOT utilize any other data or charts.



