Understanding Obesity: An Analysis of Contributing Factors

Utilizing Machine Learnng Techniques

Areti Triantafyllidou

January, 2025

Introduction

Obesity has become a critical public health issue, affecting millions worldwide. In 2022, the World Health Organization reported that 2.5 billion adults (18 years and older) were overweight, with 890 million of them living with obesity. This global problem is not just a personal health issue, it is a societal challenge that affect economies, healthcare systems, and future generations. The rapid increase in obese and overweight people over recent decades creates an urgent need to uncover its root causes and reduce its impact. Research has shown that this problem can be prevented. This report examines the factors influencing obesity by analyzing a comprehensive dataset that includes information on demographics, lifestyle behaviors, dietery habits and physical activity. Using Machine Learning techniques, this study addresses the research question: *"How effectively can machine learning models predict obesity using lifestyle data, and which factors have the greatest impact on classification?"* To answer this, the analysis applies several Machine Learning methods, including Conditional Inference Trees (CITs) Random Forests (RFs) and Support Vector Machines (SVMs). To enhance interpretability, tools such as feature importance analysis, partial dependence plots (PDPs), and accumulated local effects (ALE) plots are employed. The goal is to better understand the relationships between these factors and uncover the behavioral and dietary patterns that play the most critical role in determining an individual's risk of obesity. By using these methods, this approach provides valuable insights into the predictors of a health problem such as obesity, aiding in the development of effective prevention strategies and contributing to improved public health outcomes.

Data

The dataset utilized in this study was obtained from a publicly available dataset on Kaggle, titled "Obesity Dataset". The data was collected through survey questionnaires aimed at individuals from various age groups and backgrounds. It consists of 1610 observations and 15 variables, providing a comprehensive overview of individuals' lifestyle habits, dietary behaviors, and other health-related factors. Key attributes include demographic features such as gender, age and height, lifestyle behaviors such as dietary patterns and physical activity levels, and a target outcome classifying individuals into obesity categories: Underweight, Normal, Overweight, and Obesity. The dataset contains both numeric and categorical variables. Numeric variables include Height and Age while categorical capture information about gender, lifestyle and behaviors, such as dietary habits (e.g., frequency of vegetable intake, fast food consumption, number of meals per day, snacking), physical activity levels, and other factors like smoking habits, water intake and Tech time. Each categorical variable is encoded with specific levels to reflect distinct behaviors. For example the variable that describes the Vegetable Intake is categorized into "Rarely," "Sometimes," and "Always". The dataset underwent preprocessing steps to ensure its suitability for machine learning analysis. First of all, to address the research question effectively, the target variable was converted from its initial four categories, into a binary classification: "No Obesity" (combining "Underweight" and "Normal") and "Overweight/Obese" (combining "Overweight" and "Obesity"). This simplification helps the analysis to focus more on the factors contributing to higher overweight and obesity risks. Additionally, the "Height" variable was excluded from the analysis as it was deemed irrelevant for exploring patterns related to obesity. The "Transport" variable was restructured

into two categories: "Active" (e.g., walking, biking) and "Passive" (e.g., motorbike, automobile, public transportation). This adjustment was made to simplify the analysis and highlight the distinction between active and sedentary transportation modes. Moreover, exploratory data analysis (EDA) was conducted to examine variable distributions and relationships, providing initial insights into potential patterns and trends. Next, categorical variables were transformed into factors with meaningful labels for better interpretability and visualization. The dataset was also checked for missing values and class imbalance. No missing data were found, and the target variable displayed a balanced distribution, deleting the need for data imputation or resampling techniques. Finally, the dataset was split into training and test sets, with 70% allocated for training and 30% for testing to prepare for machine learning modeling and evaluate models' performances. These preprocessing and exploratory steps ensured a deeper understanding of the data while preparing it for effective machine learning analysis.

Methods

This study applies a combination of machine learning models to classify obesity and analyze the factors that contribute to it. Each method was selected based on its strengths in predictive accuracy, interpretability, and ability to handle the dataset's complexity. The methodology includes cross-validation, hyperparameter tuning, and interpretability techniques to ensure robust and actionable results.

Conditional Inference Tree

Conditional Inference Trees (CIT) are a type of decision tree model designed to address biases in standard decision tree algorithms. Unlike traditional trees, which tend to favor features with many split points or high variability, CIT uses statistical hypothesis testing to select splits. This process involves permutation tests that evaluate the association between predictors and the target variable, ensuring splits are based on statistical significance rather than arbitrary criteria. This approach makes CIT particularly effective for datasets with mixed types of features, such as categorical and numeric variables, and ensures an unbiased selection of features. CIT models are inherently interpretable, as they provide a visual representation of decision rules and thresholds. This simplicity allows users to easily understand the factors driving predictions, making CIT a preferred choice when interpretability is paramount. In this study, CIT was selected as the baseline model due to its straightforward interpretability and statistical rigor, providing foundational insights into the primary predictors of obesity. To optimize the model, hyperparameters were tuned through cross-validation. Cross-validation involves dividing the data into subsets (folds) and iteratively training and testing the model on these folds. This method ensures that the evaluation metrics reflect the model's performance on unseen data, reducing the risk of overfitting or underfitting.

Random Forest

The Random Forest (RF) algorithm is a powerful ensemble learning method that combines the predictions of multiple decision trees to enhance predictive accuracy and robustness. It operates through bagging (bootstrap aggregating), where each tree is trained on a unique subset of the data sampled with replacement. This approach reduces overfitting by introducing variability among trees. At each node, RF selects a random subset of features (mtry) to determine the best split, further decorrelating the trees and ensuring no single predictor dominates the model. One of RF's significant strengths lies in its ability to provide feature importance metrics, such as Gini importance or permutation importance. These measures identify key predictors and their contributions to the model, offering insights into the factors driving classifications. In this study, the mtry parameter was tuned using cross-validation to optimize the trade-off between bias and variance. The RF model also served as the foundation for global interpretation techniques, as described below.

Support Vector Machine

Support Vector Machines (SVMs) are supervised learning models designed for binary classification tasks, excelling in their ability to capture complex, non-linear relationships. The core idea of SVM is to identify the hyperplane that separates classes with the maximum margin, which enhances the generalizability of the

model. For data that are not linearly separable, SVM employs kernel functions, such as the radial basis function (RBF), to map data into higher-dimensional spaces where a hyperplane can effectively separate the classes. SVM performance is driven by two key hyperparameters: the cost parameter, C, which controls the trade-off between margin width and misclassification, and kernel width σ, which determines the influence of individual data points in the RBF kernel. Proper tuning of these parameters ensures a balance between model complexity and generalization, minimizing overfitting. In this study, SVM was employed to capture subtle interactions and non-linear dependencies in obesity classification, leveraging its robust classification capabilities.

Interpretability tools

Interpretation of machine learning models, particularly black-box methods like Random Forests, is crucial for understanding the relationships between predictors and outcomes. Global interpretability focuses on explaining the overall behavior of the model across the entire dataset, helping identify key predictors and their relationships with the target variable. Global interpretability was achieved across the models using a range of explanation tools, including feature importance analysis. This method measures the decrease in model accuracy when the values of a feature are shuffled, providing an estimate of how critical each feature is for accurate predictions. Addittionally, partial dependence plots (PDPs) was used as they illustrate how a specific feature impacts the predicted probability of obesity, averaged across all other features. Accumulated local effects (ALE) plots, extend PDPs by accounting for interactions between features, offering a more accurate depiction of how individual predictors influence model predictions. These tools provide a deeper understanding of how individual features and their interactions influence obesity classification.

The methodology begins with an interpretable baseline (CIT) and progresses to advanced models (SVM and RF) for higher accuracy. Cross-validation ensures robust evaluation, while interpretability techniques provide insights into patterns. By employing this multi-method approach, the study ensures that both the prediction task and the interpretability requirements are addressed, providing actionable insights into the factors contributing to obesity.

Analysis and Results

Conditional Inference Tree Analysis

The analysis starts with the CIT model as a baseline for predicting the obesity risk and captures the most significant variables. To optimize the model a 5-fold cross-validation was trained to ensure reliable evaluation and hyperparameter tuning. During this process, the minimum criterion for split selection (mincriterion) was tuned within a range of 0.95 to 0.99, with the optimal value determined to be 0.99. To control tree complexity and interpretability, tree depths from 1 to 10 were explored. The optimal maximum depth was determined to be 5, achieving a test set accuracy of 84.85%. However, to enhance interpretability, a depth of 4 was chosen for the final model as a shallower tree provides a clearer structure, making it easier to identify and communicate the key factors influencing obesity classification. Despite this reduction in complexity, the model maintained robust performance with the overall accuracy on the test set to reach 84.65%, indicating the proportion of correctly classified instances. It showed also a balanced sensitivity, reflecting the model's ability to correctly identify non-obese individuals, and specificity, the ability to identify overweight or obese individuals accurately, values of 85.44% and 83.27%, respectively. These metrics validate the model's ability to classify obesity while maintaining consistency across the two categories.

According to the CIT tree that it is showed below, Age was the most significant factor in classifying individuals into "Non-Obese" or "Overweight/Obese." The tree split individuals based on whether their age was above or below key thresholds (e.g., 32 years), reflecting the strong association between age and obesity risk. This highlights the strong association between aging and increased obesity risk, possibly due to metabolic changes or lifestyle factors. Following age, dietary habits such as Vegetable Intake, the number of meals per day and the consumption of fast food emerged as critical factors. This finding reflect the benefits of structured eating patterns in weight management. Gender differences were evident, with males and females exhibiting distinct risk patterns. For example, males were more likely to fall into the obese category under certain conditions. An

interesting observation from the CIT model is that higher levels of physical activity appear to be associated with an increase in obesity rates. This could reflect reverse causality, where individuals who are already obese engage in exercise to lose weight. Alternatively, it could highlight the limitations of self-reported data or suggest that exercise alone, without dietary changes, may not significantly impact weight. This result emphasizes the importance of considering multifactorial approaches to obesity management, integrating dietary, behavioral, and physiological factors.

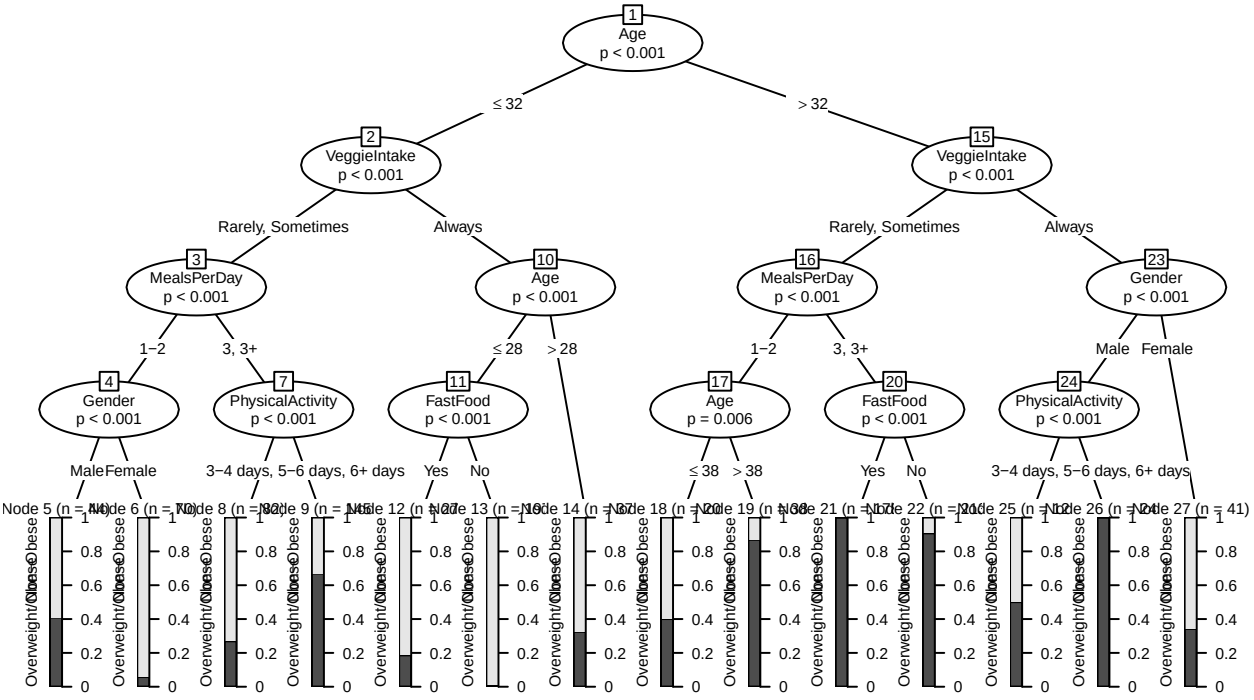

Figure 1: Conditional Inference Tree for Obesity Classification

Random Forest Analysis

Building upon the insights gained from the CIT model, the analysis moves to the Random Forest algorithm. While the CIT model provided valuable interpretability and highlighted the most influential factors affecting obesity, it is limited in capturing complex interactions between predictors. To address this, Random Forest, with its ensemble learning approach, was chosen as the next step due to its ability to handle high-dimensional data, capture non-linear relationships and improve predictive performance.

The Random Forest model was trained using 5-fold cross-validation to ensure robust evaluation and generalization to unseen data. The hyperparameter tuning process focused on the mtry parameter, which controls the number of predictors considered at each split. A grid search determined the optimal mtry value to be 6, achieving a balance between underfitting and overfitting. After training the model, its performance was evaluated on the test dataset. The model demonstrated balanced performance, excelling in both sensitivity and specificity. The confusion matrix revealed an overall accuracy at 89.83%, Sensitivity at 89.50% and Specificity at 90.11%. These results highlight the superior predictive capability of the Random Forest model compared to the CIT model.

The Random Forest model highlighted Age as the most influential predictor of obesity, with an importance score of 81.71. Eating habits such as Vegetable Intake were also a key factor, emphasizing the significance of consistent vegetable consumption in reducing obesity risk. Similarly, lower fast food consumption was associated with healthier weight categories, underscoring the impact of dietary habits. Lastly, Meals Per Day demonstrated the influence of meal frequency patterns on obesity classification. Gender played an important role, suggesting that females have lower obesity risk compared to males, particularly at younger

ages. All the evidence can be captured in the plots below. A bar plot of the top five features visually underscores their relative importance, with age standing out as the most critical factor. The PDP for Age showed a strong positive correlation with obesity risk. As age increases, the probability of being classified as "Overweight/Obese" rises sharply until a plateau is reached in older age groups. This highlights the need for targeted interventions for middle-aged adults, as this group appears most vulnerable to obesity.

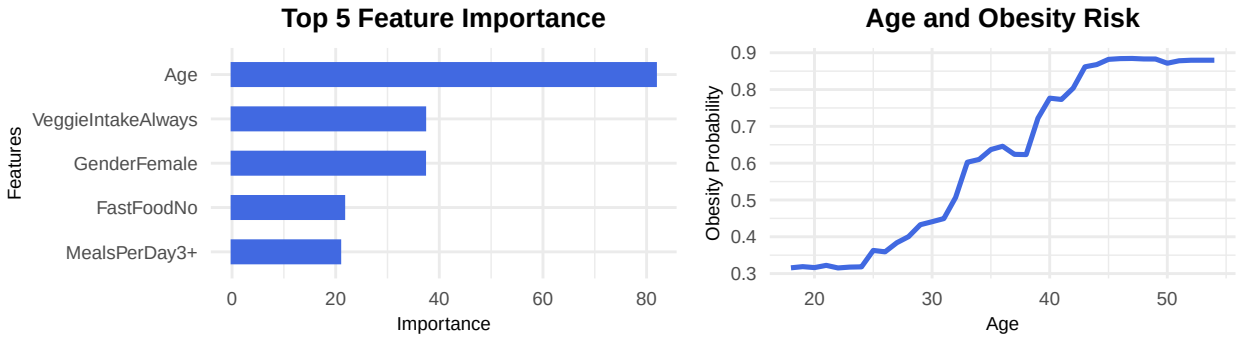

Figure 2: Top 5 Feature Importance in the Random Forest Model and Partial Dependence Plot for Age

According to RF model the strongest variables were related to demographic characteristics and eating habits. These relationships are explored through Partial Dependence Plots (PDPs), revealing insightful patterns.The Two-Way PDP for Gender and Age demonstrates how gender modifies the relationship between age and obesity. For instance, females exhibit a lower probability of obesity than males at younger ages. However, as age increases, the gap between genders narrows, suggesting that age becomes a more dominant factor influencing obesity regardless of gender. This insight suggests that interventions targeting younger males could be particularly effective. The PDP for Eating Habits explores the interaction between vegetable intake and fast food consumption. It indicates that individuals who consistently consume vegetables and limit fast food intake have a significantly lower probability of being obese. On the other hand, those who rarely consume vegetables and frequently eat fast food are at a much higher risk. Interestingly, the risk slightly decreases as vegetable intake improves from "Rarely" to "Always," indicating that balanced eating habits can somewhat mitigate the negative effects of frequent fast food consumption.This highlights the importance of dietary choices and a well-balanced diet in obesity prevention.

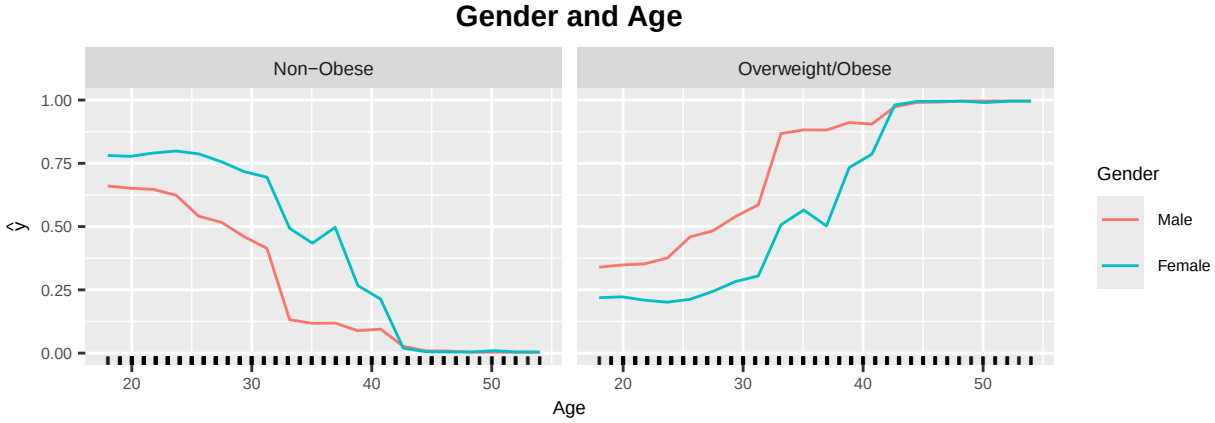

Figure 3: Partial Depedence plots for the Demographic Characteristics

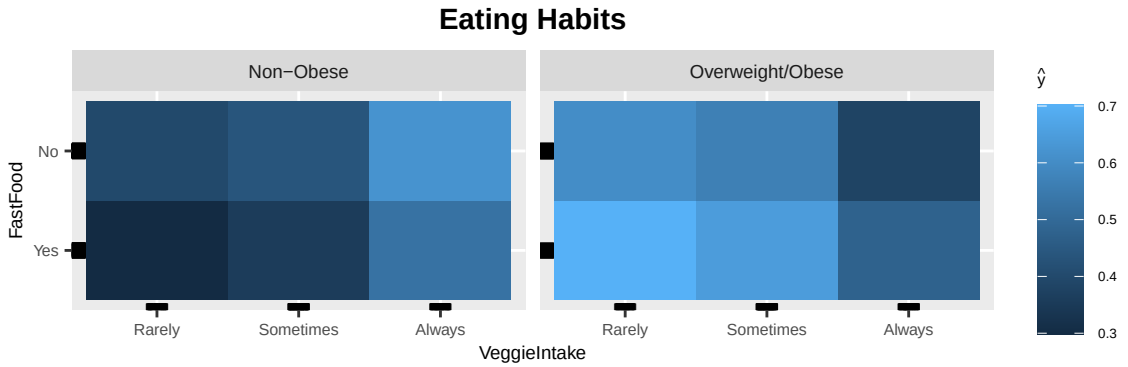

Figure 4: Partial Depedence plots for the Eating Habits

The analysis highlights that balanced dietary habits play a pivotal role in predicting obesity risk, alongside demographic factors like age and gender. To explore this further, a Partial Dependence Plot examining the interaction between snacking habits and meals per day was generated. The results reveal that frequent snacking significantly increases the probability of being classified as "Overweight/Obese" regardless of the number of meals consumed daily, emphasizing the negative impact of constant snacking. Conversely, individuals who snack "Rarely" or "Sometimes" maintain a lower obesity risk, even with a higher meal frequency (3+ meals per day). The plot also shows that frequent meals combined with constant snacking elevate obesity risk substantially, while consuming fewer meals (1–2) appears protective, even with occasional snacking. These findings underscore the importance of dietary moderation, suggesting that limiting frequent snacking and maintaining a moderate meal frequency can effectively reduce obesity risk.

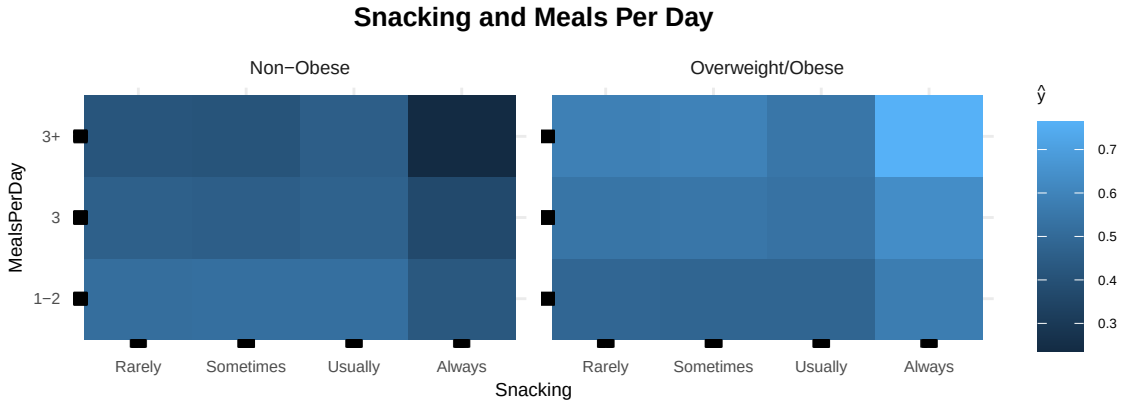

Figure 5: Partial Dependence Plot for the overall consumption of food

While the CIT model provided interpretable decision rules, such as age thresholds (e.g., below 32 years) and lifestyle habits (e.g., vegetable intake), its performance was limited by its inability to capture complex feature interactions. For instance, the Random Forest model revealed that the interaction between gender, age, and eating habits plays a critical role in obesity classification, a relationship that CIT could not fully capture.

Support Vector Machine Analysis

Next, the analysis transitioned to a Support Vector Machine model, which excels at capturing non-linear relationships through its kernel-based methods. This transition aimed to further refine the analysis and address potential limitations of tree-based models, such as over-reliance on hierarchical splits or difficulty in capturing subtle, multi-variable interactions. The radial basis function (RBF) kernel was chosen for this study, and hyperparameters, including the cost parameter C and kernel width, σ, were optimized using a

10-fold cross-validation process and the best perfoming parameters were found to be $C = 10$ and $\sigma = 0.01$. Additionally, numeric variables were scaled and centered as part of preprocessing to improve the model's efficiency and accuracy. The SVM model demonstrated strong classification results, achieving an accuracy of 86.93% and a Kappa score of 0.7357, indicating substantial agreement beyond chance. Sensitivity for the "Non-Obese" class was 84.02%, and specificity for the "Overweight/Obese" class was 89.35%, reflecting balanced performance across both classes. The balanced accuracy of 86.69% further supported its robust classification ability. A heatmap of the confusion matrix highlighted the distribution of correct and incorrect predictions, with relatively low false positive (28 cases) and false negative (35 cases) rates.

A key insight from the SVM model was the ranking of feature importance, calculated using cross-entropy loss. Age remained the most significant predictor, consistent with prior analyses. Veggie Intake also ranked highly, underscoring its positive association with healthier weight outcomes. Interestingly, Physical Activity gained greater prominence in the SVM analysis. This may suggest that SVM better captures interactions between activity levels and other predictors, such as dietary habits. Conversely, Fast Food, which was significant in the Random Forest model, ranked lower in SVM. This discrepancy might arise from overlapping contributions with other variables or the SVM's preference for non-linear feature prioritization. Features like Gender, Meals Per Day, and Snacking consistently contributed across models, although their relative importance varied, reflecting the distinct optimization strategies of SVM and Random Forest. In conclusion, the SVM model offered complementary insights, refining the understanding of feature interactions and confirming the critical role of demographic and dietary patterns in obesity classification. The differences between the models' interpretations emphasize the importance of leveraging multiple algorithms for a comprehensive analysis.

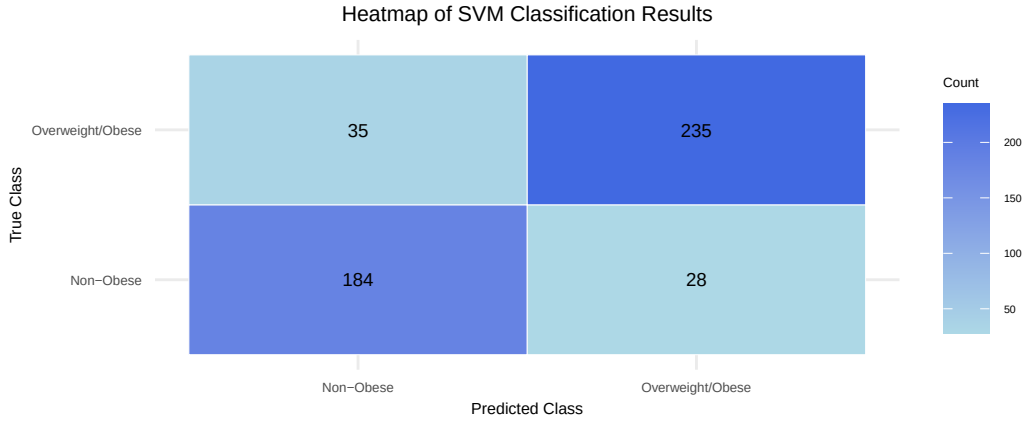

Figure 6: Heatmap for models' metrics

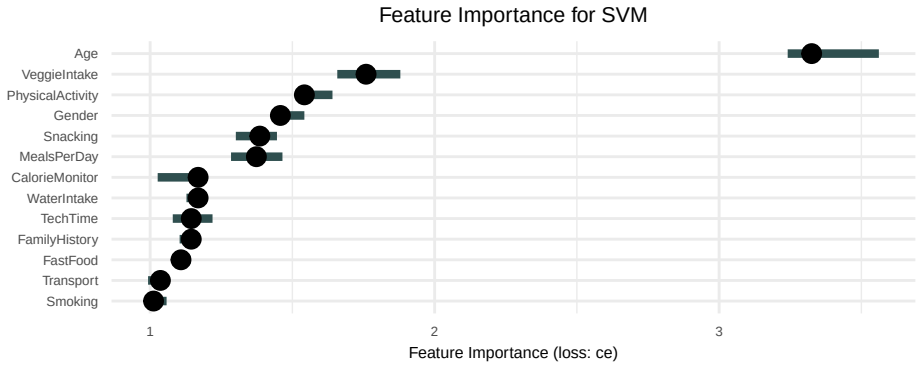

Figure 7: Feature Importance of SVM model

To further the insights, a variable of the models with the highest acuracy, Random Forest and SVM was

selected for comparison. Both models highlighted the importance of healthy eating habits, with vegetable consumption emerging as one of the most critical factors. The ALE plots below illustrate the similarities and differences in the predictions of the two models regarding Veggie Intake. Both models agree that the "Always" category is strongly associated with "Non-Obese", while the "Rarely" and "Sometimes" categories are more associated with "Overweight/Obese". However, differences in the intensity of the effect are evident. Random Forest shows greater differentiation across categories, suggesting it relies more heavily on this variable. SVM, on the other hand, displays more balanced effects, potentially reflecting its ability to handle non-linear relationships in the data. These differences highlight how the two models interpret the same variable in distinct ways. Beyond dietary factors, demographic characteristics such as Age and Gender consistently emerged as significant predictors in all models. The Conditional Inference Tree (CIT) model, which served as a baseline, provided interpretable insights into key variables but lacked the capacity to capture complex interactions. Random Forest improved upon this by aggregating decisions across multiple trees, capturing higher-order interactions while retaining interpretability through feature importance rankings. SVM further advanced the analysis by modeling subtle non-linear relationships that are not easily captured by tree-based methods.

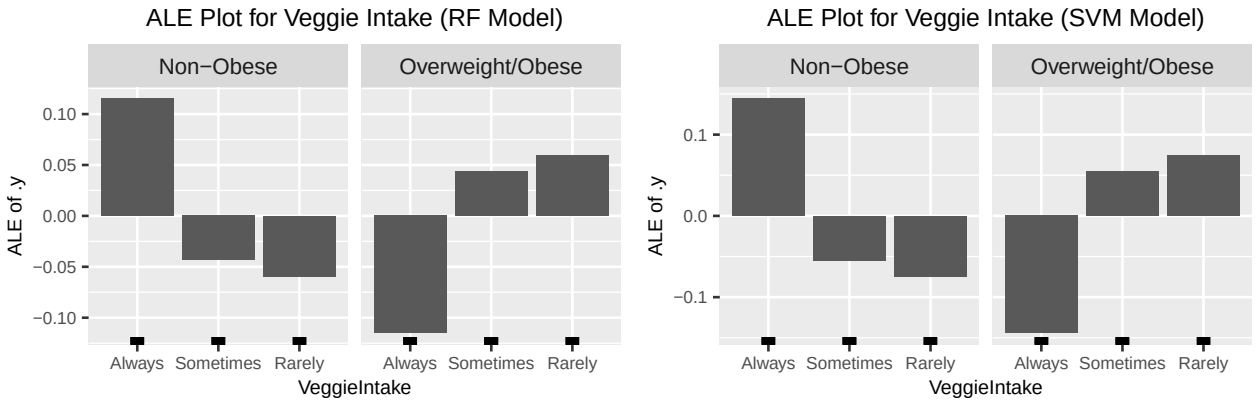

Figure 8: ALE plots for VeggieIntake for RF and SVM models

The model comparison highlights that the Random Forest algorithm outperforms the other methods across all metrics, achieving the highest accuracy (89.83%), sensitivity (89.50%), specificity (90.11%), and kappa (79.52%). This reflects its ability to handle complex relationships and provide robust predictions. The Support Vector Machine (SVM) closely follows with an accuracy of 86.93% and a balanced performance in sensitivity (84.02%) and specificity (89.35%), indicating its effectiveness in capturing non-linear relationships. Meanwhile, the Conditional Inference Tree (CIT) demonstrates interpretability with an accuracy of 84.44%, but it falls slightly short compared to the other models, especially in specificity (83.27%) and kappa (68.77%). Overall, while Random Forest provides the best performance for predictive accuracy, CIT offers greater simplicity, making it suitable for scenarios requiring model transparency. The choice of the most suitable model depends on the study's goals. For clear and interpretable insights, the CIT model is ideal, while Random Forest balances interpretability and predictive power. SVM model, on the other hand, excels at capturing complex, non-linear interactions, with high predictive accuracy.

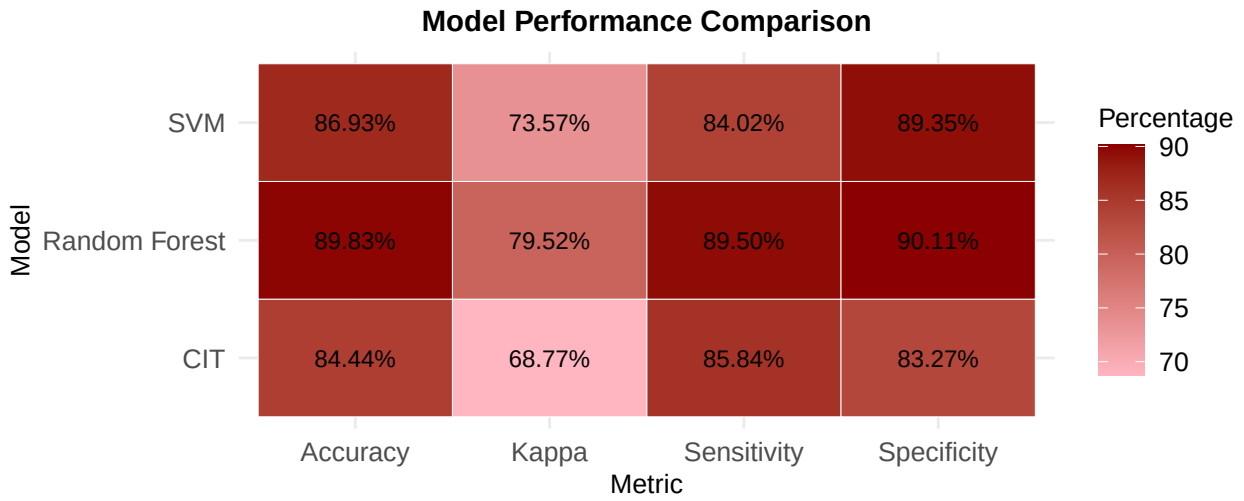

Figure 9: Heatmap for models' metrics

Conclusions

The findings of this study address the research question *"How effectively can machine learning models predict obesity using lifestyle data, and which factors have the greatest impact on classification?"* by using a multi-method approach, including Conditional Inference Trees (CIT), Random Forest (RF) and Support Vector Machines (SVM), with a combination of global interpretability methods. The study demonstrates that machine learning models can effectively classify obesity and shows the most important characteristics.The Random Forest model showed the highest accuracy, with the SVM model closely following, while the CIT model achieved a slightly lower accuracy. Across all models, Age was the most significant predictor, with older individuals facing a higher risk of obesity. Dietary habits, particularly Vegetable Intake, consistently ranked as critical factors, alongside lifestyle behaviors such as physical activity and fast food consumption. Gender differences were also observed, with females generally exhibiting a lower obesity risk at younger ages. To address obesity, strategies to reduce obesity should prioritize promoting healthy eating behaviors, such as consistent vegetable consumption, limiting fast food intake, and maintaining balanced meal frequencies. Public health campaigns can focus on raising awareness about the benefits of it and provide incentives for healthier food options in schools, workplaces, and communities. As age showed the dominant factor, older adults might focus on managing existing risk factors through regular health screenings, personalized nutrition plans, and community-based fitness programs. While this study provides valuable insights, it has certain limitations. The dataset did not include socioeconomic or psychological variables, which are known to influence obesity. Including these factors could enhance model performance and offer a more comprehensive understanding of obesity determinants. Future research could explore these interactions more deeply to develop age-specific and personalized health strategies. The use of longitudinal data could help capture temporal patterns, improving the ability to predict obesity trends over time.

References

1. Breiman, L. (2001). "Random Forests". Machine Learning, 45(1), 5-32.
2. Cortes, C., & Vapnik, V. (1995). "Support-vector networks". Machine Learning, 20(3), 273-297.
3. Suleyman Sulak, "Obesity Dataset", Kaggle. Available at: https://www.kaggle.com/datasets/suleymansulak/obesity-dataset.
4. World Health Organization (WHO). (2021). "Obesity and overweight". Available at: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.
5. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

6. Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Available at: https://christophm.github.io/interpretable-ml-book/
7. Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychological Methods, 14(4), 323–348.