Introduction

Obesity is a significant health issue that affects millions of people worldwide, leading to various chronic conditions such as diabetes, cardiovascular diseases, and other metabolic disorders. This project aims to estimate obesity levels based on eating habits and physical condition, addressing a critical public health concern. By predicting obesity levels accurately, we can contribute to better health outcomes by identifying at-risk individuals early and suggesting appropriate interventions.

Leveraging machine learning techniques, this project analyzes and predicts complex health-related conditions, demonstrating the practical application of data science in the healthcare sector. The use of a rich dataset, encompassing various attributes related to lifestyle and physical conditions, makes this an intriguing and challenging problem. A good predictive model for obesity levels is crucial as it aids in the early detection and prevention of obesity-related health issues, improving the quality of life and reducing healthcare costs. Additionally, this model can assist public health officials in designing better health policies and programs tailored to specific populations, thereby having a broader impact on public health management.

Methods Section

Data Exploration

The first step in our analysis involved exploring the dataset to understand the distribution and characteristics of the data. We used various statistical measures and visualizations to summarize the dataset. Key features such as age, height, weight, and other attributes were analyzed. We generated histograms to observe the distribution of numerical variables and bar charts for categorical variables. Additionally, correlation matrices were created to identify relationships between different features, helping us understand how variables are interrelated.

Preprocessing

We did the preprocessing prepare for the data for model training. The initial dataset included both categorical and numerical features, as there were none missing values, we directly began the encoding. Categorical variables such as gender, family history with overweight, frequent consumption of high caloric food (FAVC), smoking habits, and monitoring of calories consumed (SCC) were encoded using label encoding. Variables like eating frequency between meals (CAEC), alcohol consumption (CALC), and mode of transportation (MTRANS) were one-hot encoded. Numerical features were normalized using MinMaxScaler for logistic regression and standardized using StandardScaler for the decision tree model to ensure uniformity in the data scale.

Model 1: Logistic Regression

The logistic regression model was employed to classify obesity levels. The data was split into training and test sets with a test size of 10%. The logistic regression model was trained using the normalized dataset, with the maximum iteration set to 1000 to ensure convergence. We used the ‘lbfgs’ solver and multinomial loss for handling multi-class classification. The model's performance was evaluated using accuracy scores on both training and test sets, and learning curves were plotted to visualize the model's learning process over different training set sizes.

Model 2: Decision Tree

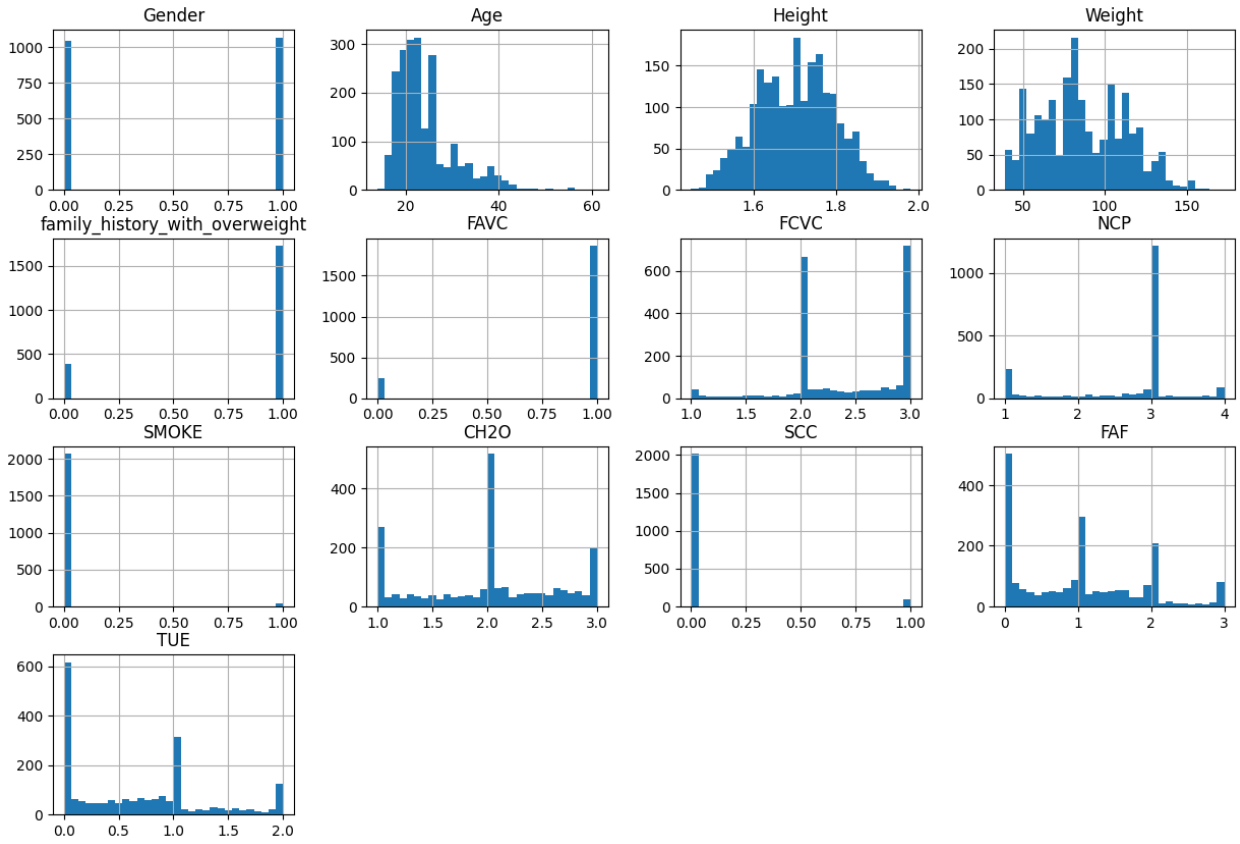
A decision tree classifier was chosen as the second model. The data was split into training and test sets with an 80-20 split. The decision tree model was trained using the standardized dataset. We set the random state to 42 for reproducibility. The model's performance was evaluated using accuracy scores, and confusion matrices were generated to understand the classification results in detail. Additionally, hyperparameter tuning was performed using GridSearchCV to optimize the model. The parameters tuned included criterion (gini, entropy), max_depth (None, 10, 20, 30, 40, 50), min_samples_split (2, 5, 10), min_samples_leaf (1, 2, 4), and

max_features (None, sqrt, log2). The best parameters and the best score obtained from grid search were recorded.

Results Section

Data Exploration

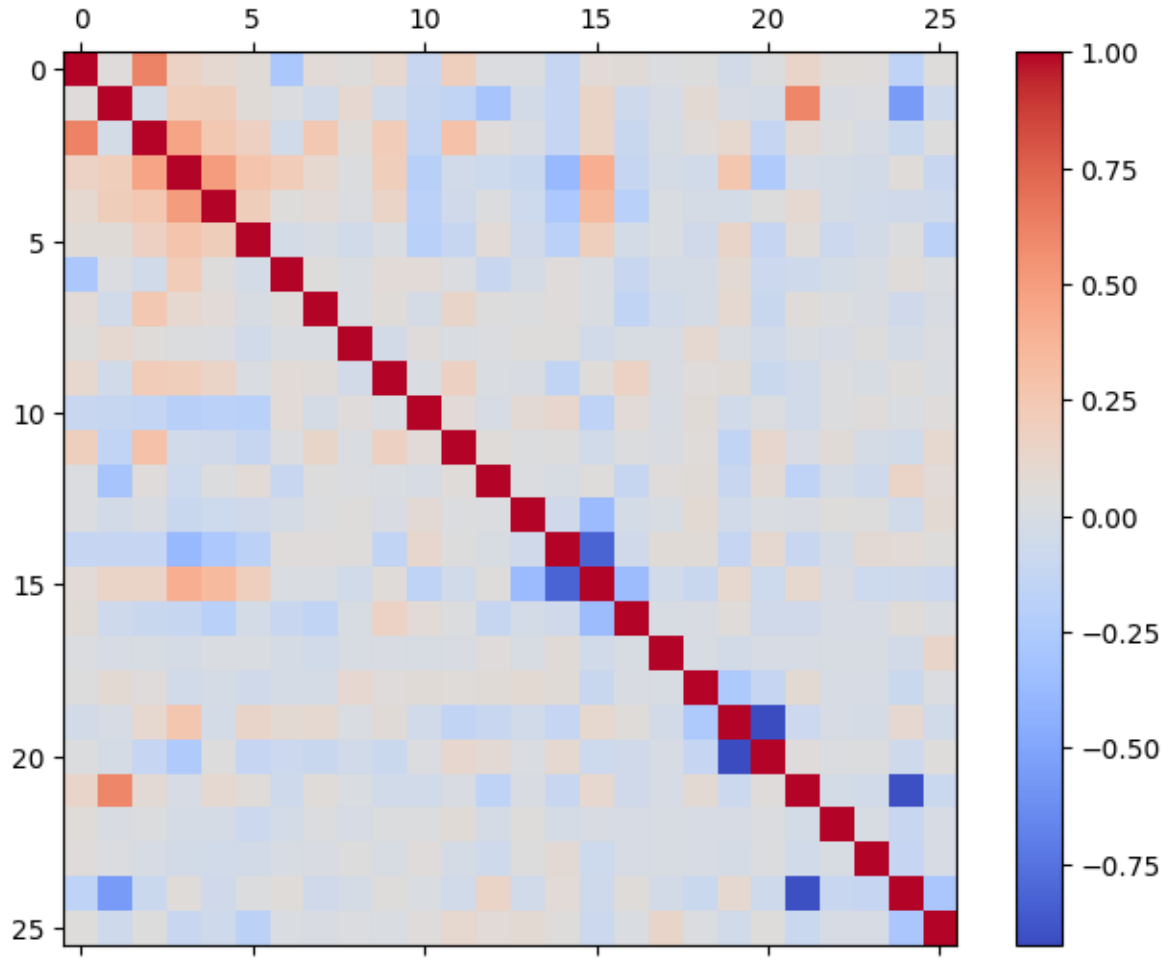
The data exploration phase revealed insights into the dataset's distribution and characteristics.



Histograms

Histograms showed that the age, height, and weight variables were reasonably well-distributed with no significant skewness. Bar charts of categorical variables, such as family history with overweight and eating

habits, provided a clear view of the frequency of different categories.



Correlation Matrix

The correlation matrix indicated that certain features, such as weight and height, had stronger relationships with the target variable, NObeyesdad (obesity levels).

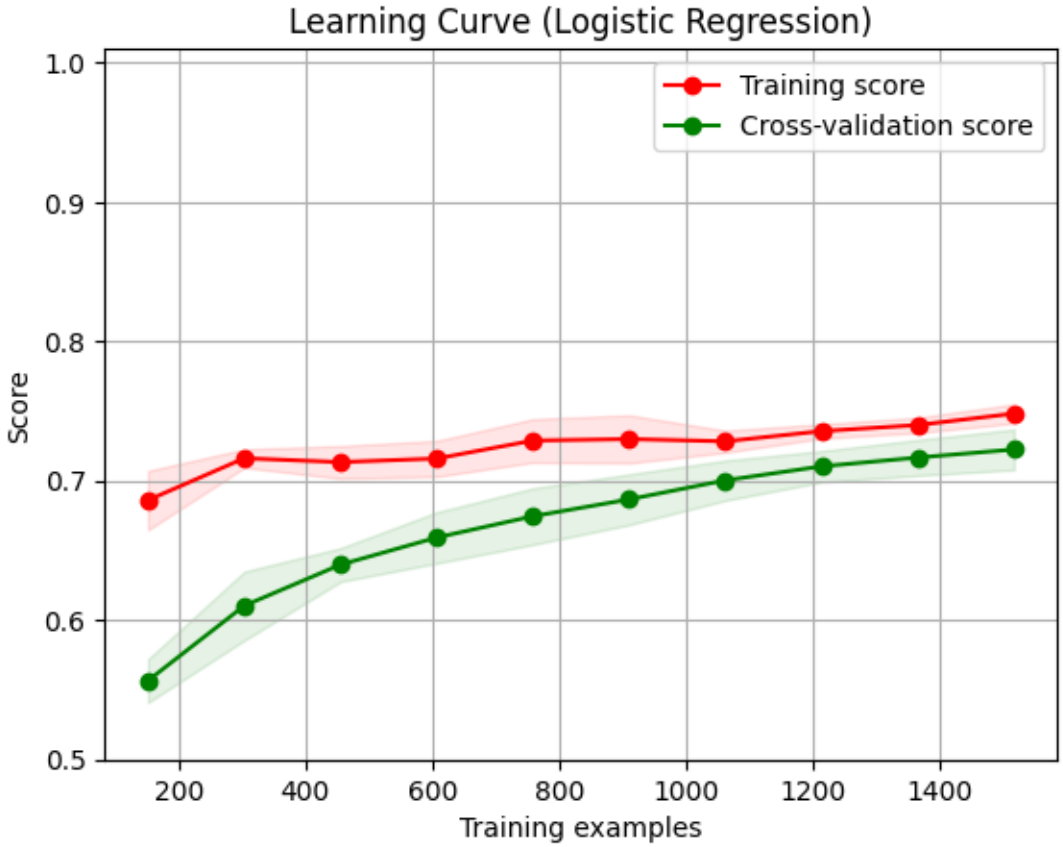
Preprocessing

Preprocessing transformed the dataset into a suitable format for model training. The label encoding and one-hot encoding steps were successful, converting categorical variables into numerical formats. Normalization using MinMaxScaler brought all features into the [0, 1] range, while standardization using StandardScaler adjusted the mean and variance of numerical features. These preprocessing steps ensured that the logistic

regression and decision tree models could operate effectively on the transformed data.

Model 1: Logistic Regression

The logistic regression model, trained on the normalized dataset, achieved a training accuracy of 75.88% and a test accuracy of 71.70%.

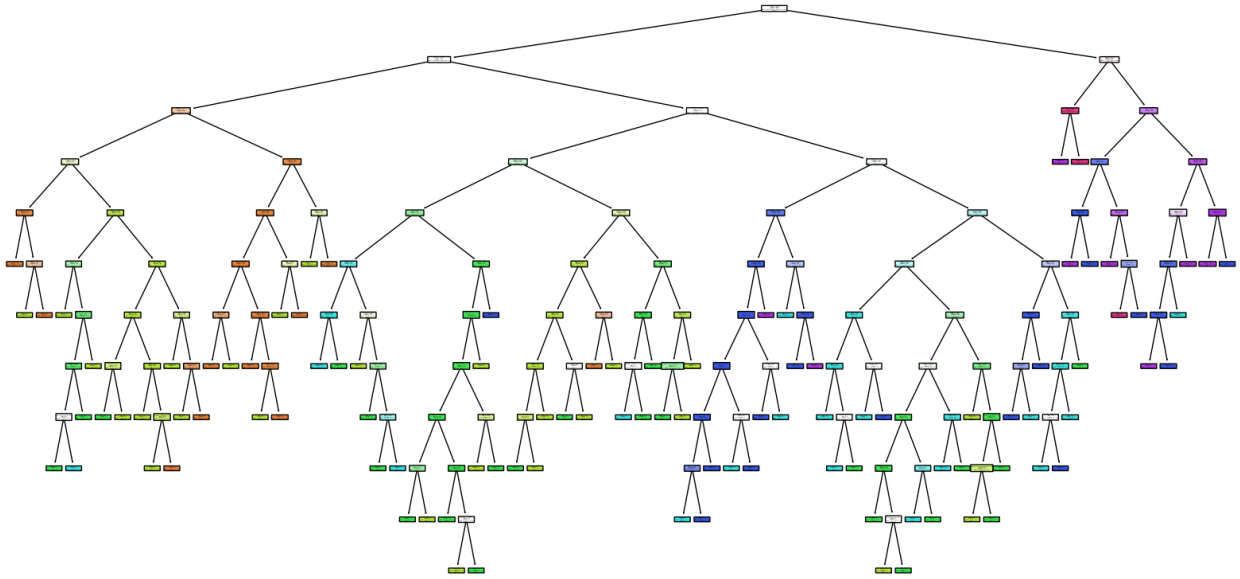


The learning curve for logistic regression showed a consistent improvement in accuracy with an increase in the training data size, indicating that the model was learning effectively. The model's predictions were generally accurate, although there were some misclassifications, particularly in distinguishing between adjacent obesity levels.

Model 2: Decision Tree

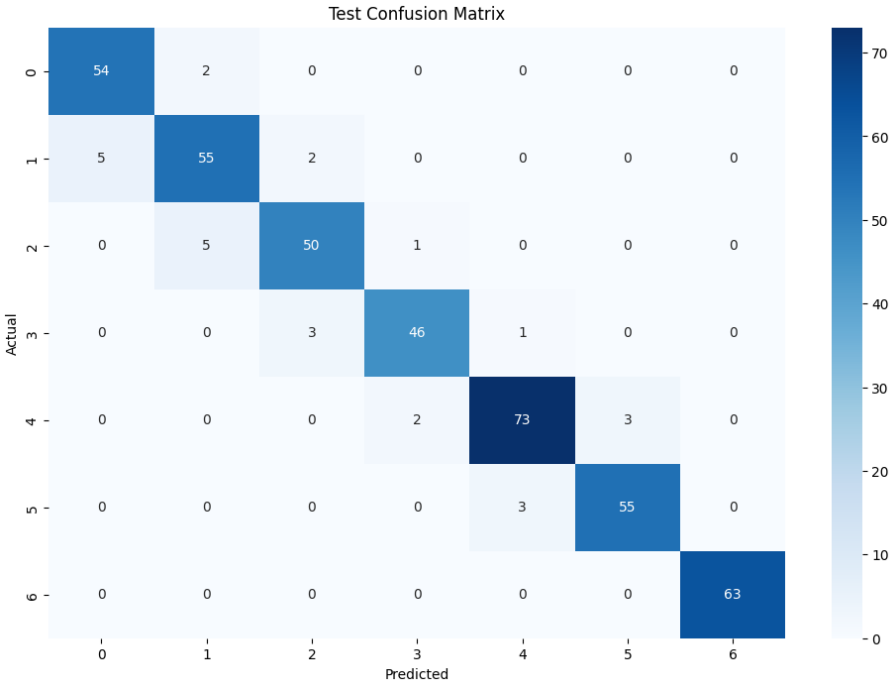
The decision tree classifier, trained on the standardized dataset, performed exceptionally well, achieving a training accuracy of 100% and a test

accuracy of 93.62%.



Decision Tree

The confusion matrix for the test set demonstrated that the model accurately classified most of the obesity levels, with very few misclassifications. Hyperparameter tuning further optimized the model, and the best parameters found included using the 'entropy' criterion and no maximum depth, among others. These parameters resulted in an optimized model with an improved classification performance.



Discussion Section

Data Exploration

We first dived into the data with basic visualizations, like histograms and bar charts, to get a feel for the dataset. These visualizations gave us insights into how age, height, and weight were distributed across our samples. In doing that, it helps us identify any imbalances or skewness that might affect our model's performance, in aspects such as weight calculation. Therefore, if the data is not well-distributed, we might need to rescale it to ensure that each feature contributes equally to the model. Similarly, we also examined categorical variables, such as family history with overweight and eating habits, to see their spread. Besides that, the correlation matrix was particularly useful for it highlighting which features were more closely related to obesity levels. It shall guide our feature selection and engineering process, though we quickly realized the complexity of predicting obesity with so many intertwined factors.

Preprocessing

Next, we transform our raw data into something our models could digest. This involved encoding categorical variables so they could be treated as numbers by our models. We used label encoding for binary categories and one-hot encoding for others like eating habits and transportation modes. In this project, normalizing and standardizing the data was a crucial step to ensure that features with different scales didn't mess up our models. Normalization was especially important for logistic regression, as it ensures that no single feature dominates the model due to its scale, while standardization helped with the gradient descent calculation. However, dealing with the multi-class nature of our target variable added an extra layer of complexity.

Model 1: Logistic Regression

We started with logistic regression because it’s straightforward and often serves as a reliable baseline model for classification problems. We also value its ability to provide interpretable coefficients, which might be helpful for understanding the impact of each feature on the prediction. It’s also computationally efficient, making it suitable for initial experiments and quick iterations.

Turns out that training the logistic regression model yielded a training accuracy of around 75.88% and a test accuracy of 71.70%. These results indicate that the model performed not too well in identifying obesity levels based on the given features. And the learning curve showed that even though the model's performance improved with more data, additional data that enhance its accuracy is needed.

Therefore, it highlighted the limitations of logistic regression. As a linear model, logistic regression assumes a linear relationship between the input features and the target variable. This assumption makes it less effective at capturing the complex, non-linear relationships inherent in our dataset. Overall speaking, the logistic regression model's linear approach was a bit too simplistic for our complex problem.

Model 2: Decision Tree

Switching to the decision tree model was like turning on a high-powered microscope. The decision tree could capture the non-linear relationships much better, which was reflected in its higher accuracy—100% on training and 93.62% on testing. The high training accuracy was a bit concerning since it hinted at overfitting, but the test accuracy showed the model was still generalizing well. Hyperparameter tuning helped us refine the model further. We tried different criteria like 'gini' and 'entropy', and varied the max depth and other parameters, which improved the model's performance. Visualizing the decision tree was enlightening because it showed us which features were most influential in predicting obesity levels.

Believability and Robustness of Results

Overall, our results seem pretty believable. The logistic regression model provided a good baseline but had its limitations with non-linear data. The decision tree model performed much better, though we had to be cautious about overfitting. Even though the test seems well, the robustness of our results shall be further tested with more data and by experimenting with other models to provide a more balanced approach to bias and variance, or using some regulation methods in preventing that.

Conclusion

Reflecting on our project, we made strides in understanding and predicting obesity levels based on eating habits and physical conditions. Our journey began with thorough data exploration, followed by employed logistic regression and decision tree models, each offering unique insights and presenting distinct challenges. While logistic regression provided a solid baseline, it struggled with the complexity of our data. The decision tree model performed admirably, but its tendency to overfit reminded us of the need for careful model tuning and validation.

In hindsight, there are a few things we can be improved on. Firstly, incorporating a wider variety of model might be provided a more robust understanding of the data. These models often balance bias and variance better than single models. Additionally, exploring deep learning techniques could have uncovered more intricate patterns and interactions within our dataset. Enhancing our feature engineering process, possibly by including more domain-specific knowledge, might have also improved model performance.

Looking ahead, future directions could include developing a more integrated approach that combines multiple models to leverage their strengths. For instance, exploring advanced preprocessing techniques and

dimensionality reduction methods, such as PCA, might refine our feature set.

Collaboration

Frank Li: discussing high level ideas, data preprocessing, complementing the written report, organizing the github repository.

Po-cheng Lai: discussing high level ideas, data preprocessing, complementing the written report.

Yixuan Li: discussing high level ideas, data preprocessing, training the first model, complementing the written report.

Zhaogu Sun: discussing high level ideas, data preprocessing, training the second model, complementing the written report.