Abstract

This project examines the impact of lifestyle factors on the risk of developing obesity specifically in individuals from Mexico, Peru, and Colombia. The data set includes qualitative and quantitative attributes related to eating habits (e.g., frequent consumption of high-caloric food, vegetable consumption frequency, number of meals per day, food intake between meals, daily water intake, and alcohol consumption). The data also capture physical activity attributes (e.g., physical activity frequency, technology usage time, and transportation mode). Our objective was to determine which lifestyle characteristics contribute to predicting obesity risk and identify key factors influencing the target variable, the level of obesity risk an individual has.

Our results indicate that family history of being overweight and gender are significant positive predictors of obesity risk. On the other hand, height and physical activity frequency (FAF) are negative predictors. Taller individuals and those with higher physical activity frequency were associated with lower predicted obesity risk, assuming all other factors remained constant.

The residual analysis of the linear regression model supports its validity since the residuals appear to be normally distributed and centered at zero. However, the linear regression model failed to predict the target variable within the range of 0 through 6 which indicates that the model could be improved upon and that it would be beneficial to explore non-linear models. The classification report of the logistic regression model demonstrates strong performance for certain classes, particularly class 6, which has the highest precision and recall. The model also performs well for class 0 (normal weight) and class 5 (obese-level I). However, it struggles with

intermediate obesity classes, Overweight Level I (class 2) and Overweight Level II (class 3), where precision and recall are lower in comparison.

Introduction

Obesity is a global health challenge, contributing to increased morbidity and mortality rates from associated diseases such as diabetes, cardiovascular issues, and certain cancers. The prevalence of obesity varies by region and is influenced by complex interactions between genetic, environmental, and lifestyle factors. Our study aimed to contribute to the understanding of these factors by examining how specific lifestyle factors impact the risk of obesity in individuals in Mexico, Peru, and Colombia. This research is rooted in the idea that regionally tailored interventions can provide more effective solutions than generalized approaches. The primary objective of this project was to analyze and identify lifestyle factors that contribute most significantly to obesity risk. By understanding which dietary and physical activity patterns predict obesity levels, we aim to provide actionable insights that can inform public health strategies and individual interventions. Specifically, we sought to identify key predictors of obesity risk within the data set, including dietary habits, physical activity frequency, and other lifestyle factors, evaluate the performance of statistical and machine learning models—including linear regression and logistic regression—in classifying individuals into specific obesity risk categories, and assess the cultural and regional influences on obesity risk to provide context-specific recommendations for intervention.

Our analysis utilized a dataset derived from a learning model trained on obesity risk and cardiovascular disease data. The dataset encompasses both qualitative and quantitative variables, including dietary habits such as the frequency of high-calorie food consumption, number of daily

meals, vegetable consumption, water intake, and alcohol consumption; physical activity attributes such as the frequency of exercise, transportation mode, and time spent using technology; and demographic and physiological factors such as gender, family history of obesity, height, and weight. By leveraging these diverse variables, we aimed to capture the multifaceted nature of obesity risk.

Our findings showed the significant role of family history and gender as positive predictors of obesity risk. Individuals with a family history of being overweight and females are at higher risk of developing obesity. Conversely, taller individuals and those engaging in more frequent physical activity exhibited a lower risk. The linear regression model, while offering a high R^2 value (0.912), demonstrated limitations in predicting the full range of obesity risk categories, leading to the need to explore non-linear models. Logistic regression, on the other hand, provided strong classification performance for extreme obesity levels (e.g., Obesity Type III) but struggled with intermediate categories such as Overweight Levels I and II.

The dataset's focus on individuals from Mexico, Peru, and Colombia allowed us to analyze obesity risk within a culturally and regionally specific context. This is particularly important given that dietary patterns, levels of physical activity, and transportation modes can vary significantly across regions. For example, traditional diets and the prevalence of walking or biking as transportation modes in these countries provide unique insights into obesity risk factors that may differ from those in other regions. Understanding the predictors of obesity risk is essential for designing effective interventions. For instance, public health campaigns could prioritize promoting physical activity and providing education about healthier dietary habits tailored to the specific cultural contexts of Mexico, Peru, and Colombia. Additionally, the insights gained from this study can inform policies aimed at addressing the environmental and

socioeconomic factors that exacerbate obesity risk. In summary, our research highlights the interplay between lifestyle choices and obesity risk, offering insights into region-specific health challenges. The subsequent sections will provide a detailed overview of the dataset, methods, results, and implications for future work, outlining the strengths and limitations of our approach.

Data

Data Representation

Our data was generated from a deep learning model trained on the "Obesity or CVD Risk" dataset which estimates the obesity levels in people from Mexico, Peru, and Colombia (AravindPCoder). The data included attributes related to eating habits (frequent consumption of high-caloric food, frequency of vegetable consumption, number of main meals consumed per day, consumption of food between meals, daily water consumption, and alcohol consumption). The data also included attributes related to an individual's physical state (whether the individual monitors their caloric intake, frequency of physical activity, time using technology, and transportation mode used). Overall, the data was clean since each row in the data was a single observation corresponding to lifestyle attributes to an individual, and each column in the data represented a different variable which can only take on one value per observation.

Diet, Habits, and Obesity Risk

The data were useful for studying the phenomenon we were interested in a few different ways. The attributes related to eating habits like consumption of high-calorie food, the amount of meals consumed per day, and the consumption of food between meals gave important insight into the different types of caloric intake patterns. The caloric intake patterns can be seen as

factors that were directly linked to obesity, so understanding them helped to identify the dietary behaviors that contributed significantly to obesity risk.

Physical Activity and Obesity Risk

Variables such as the frequency of physical activity and the time spent using technology helped gauge the level of a more active versus a more sedentary lifestyle. Since physical inactivity is a major contributor to both obesity and CVD, these variables were necessary to understand how lifestyle choices affect health outcomes.

Water and Alcohol Consumption

Drinking water is crucial to function, and water consumption may have an influence on obesity and health as a whole (Khil et al. 1). On the other hand, drinking alcohol in excessive amounts is widely known to increase caloric intake and often leads to poor health outcomes, and it may be "a risk factor for obesity in some individuals (Traversy and Chaput 122). These variables show how water and alcohol consumption contribute to overall obesity risk.

Cultural and Regional Insights

Since the data we gathered was specific to populations in Mexico, Peru, and Colombia, they allowed for insights that were culturally and regionally relevant to our project. By understanding how dietary and lifestyle patterns vary across these populations it helped us to identify specific risk factors that were tied to regional habits and available resources

Transportation and Physical Health

The transportation mode used reflected physical activity, especially if people used active types of transportation like walking and biking instead of driving to travel. This variable is often

linked to obesity and CVD risk since frequently using more active modes of transportation reduces sedentary time and likely promotes healthier body weight.

One challenge we anticipated–data cleaning–was largely resolved because the data set we selected came from Kaggle, and was already cleaned for analysis. This allowed us to focus on performing EDA and creating visualizations and descriptive statistical tables. For future implications, one area of concern is whether the cultural and/or socioeconomic differences that are inherent to the data will impact the accuracy of the model if it were to be used to predict obesity risk in a different population since the data were generated from a dataset that collected information from individuals from Mexico, Peru, and Colombia. Another potential area of concern is the potential outliers within the dataset, we can see with the graphical models some extreme peaks within the BMI graph. It is evident that there are more participants in certain categories than others. This can skew our results for the prediction model and distort parameter estimates, leading to poor generalizations as well.

Fig. 1. Data Dictionary

Data Dictionary

Variable Name	Description
ID	unique identifier for each observation
Gender	gender of the individual (male or female)
Age	age of the individual (years)
Height	height of the individual (m)
Weight	weight of the individual (kg)
family_history_with_overweight	whether the individual has a family history of being overweight (yes or no)
FAVC	whether the individual frequently consumes high-caloric foods (yes/no)
FCVC	whether the individual frequently consumes vegetables (yes/no)
NCP	number of main meals eaten per day
CAEC	frequency of eating food in between meals (Sometimes, Frequently, Always)
SMOKE	whether the individual is a smoker (yes/no)
CH2O	daily water consumption
SCC	whether the indiviudal monitors the amount of calories they consume (yes/no)
FAF	frequency of physical activity per week
TUE	time spent using technology/devices (hours per day)
CALC	frequency of the individual consuming alcohol (Sometimes, Frequently, Always)
MTRANS	mode of transportation used
NObeyesdad	target categorical variable of the individual's obesity risk level

Fig. 2. Age Distribution

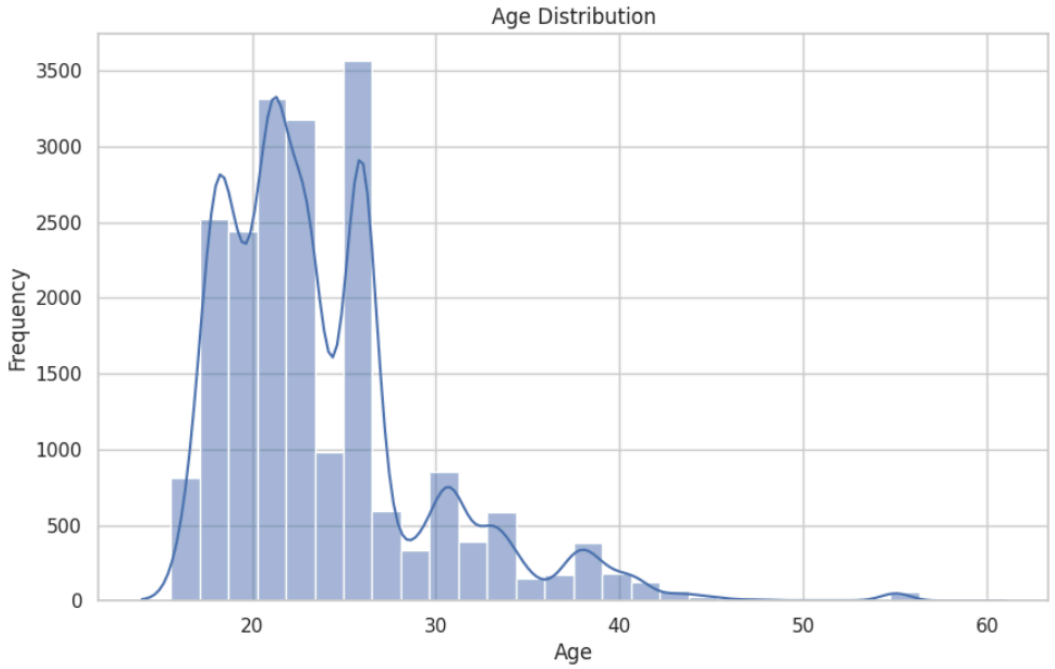


Fig. 3. Weight Distribution

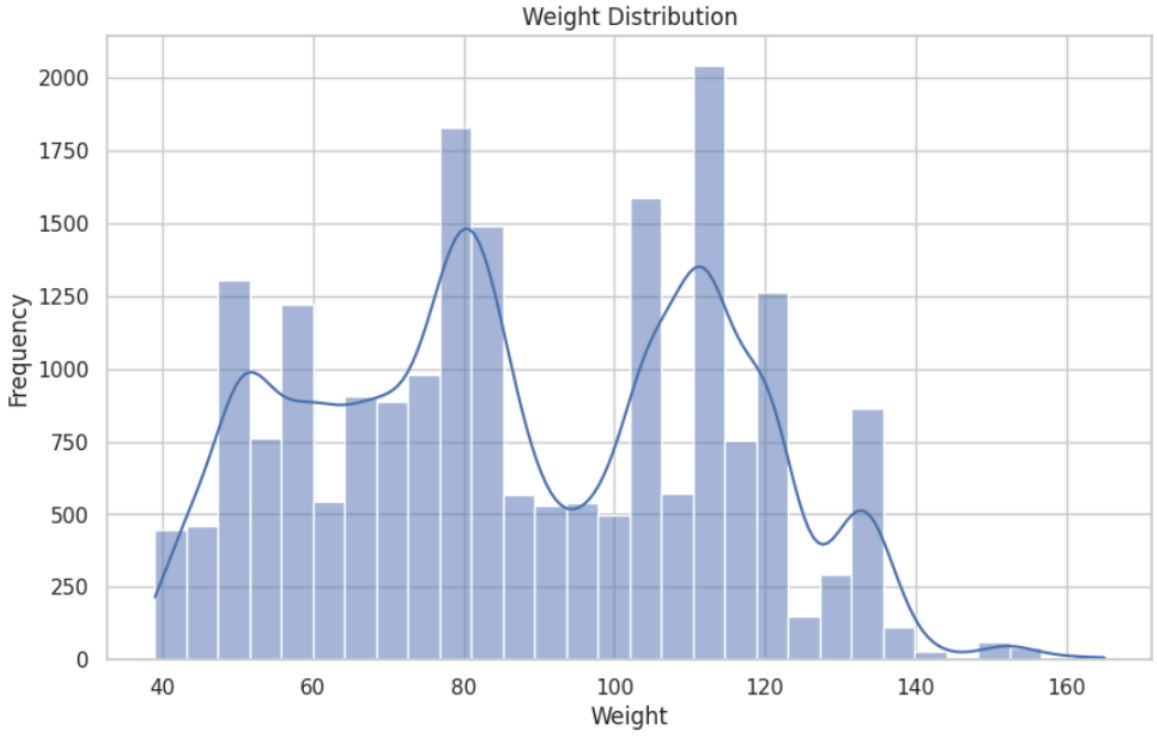


Fig. 4. Height Distribution

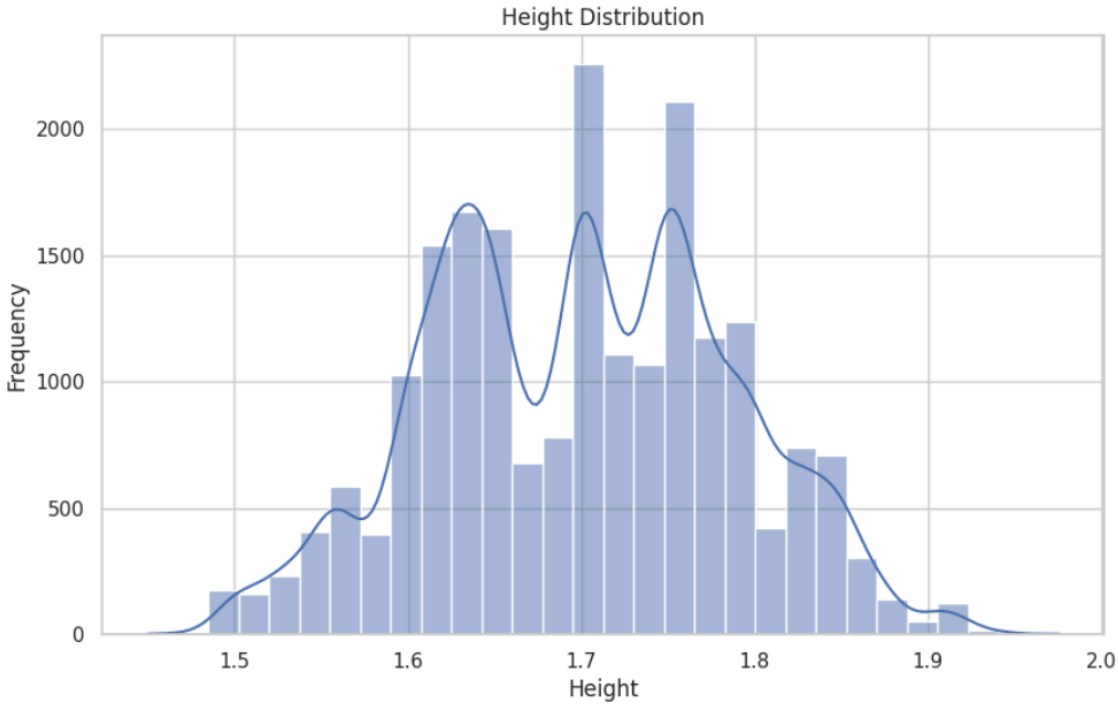


Fig. 5. Gender Distribution

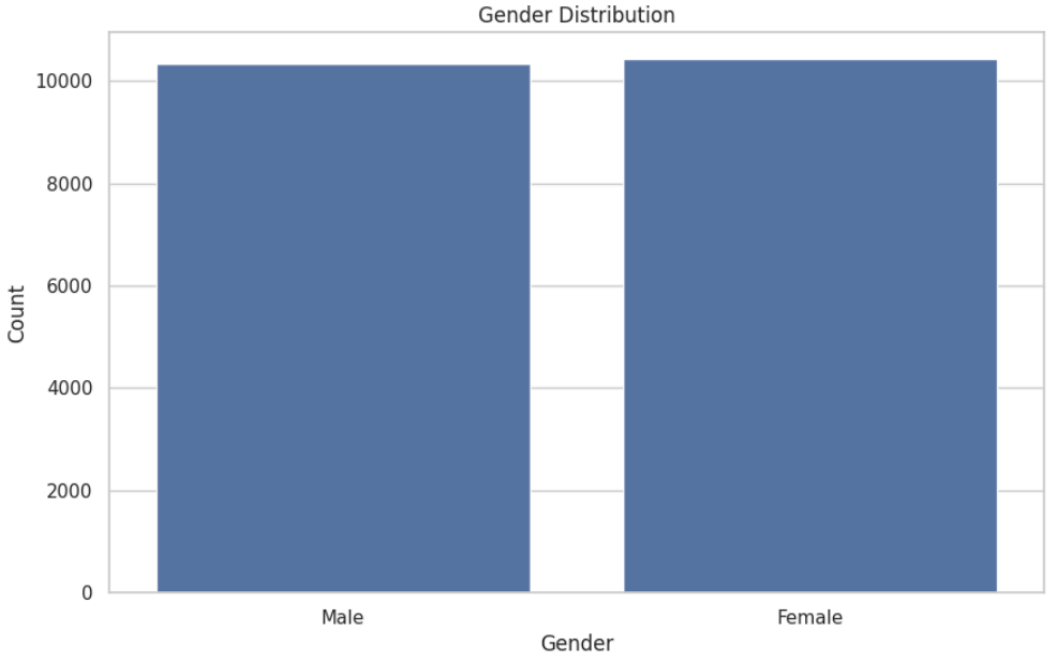


Fig. 6. Family History of Overweight

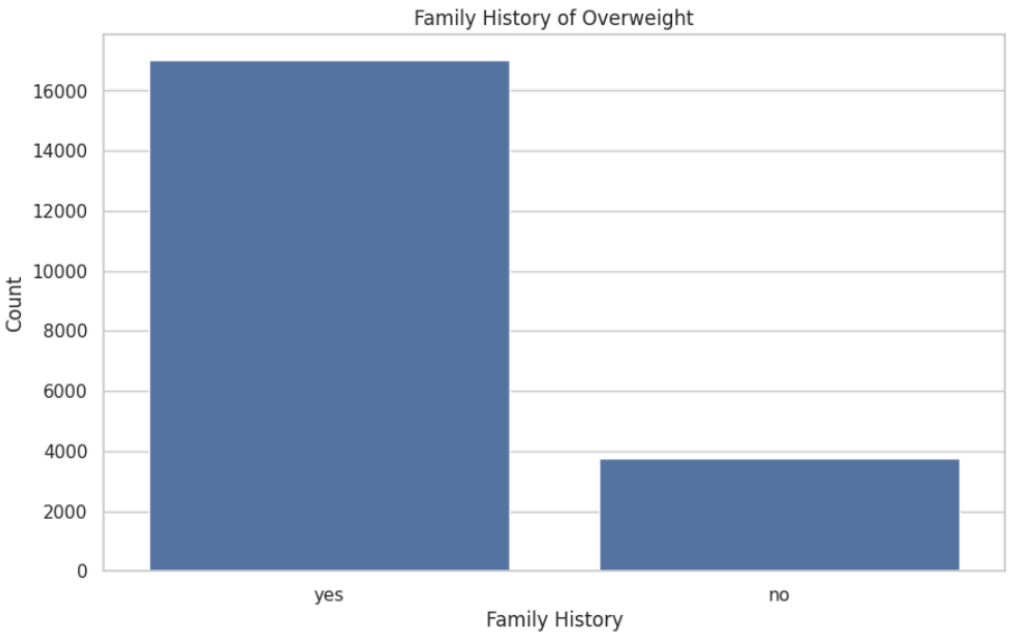
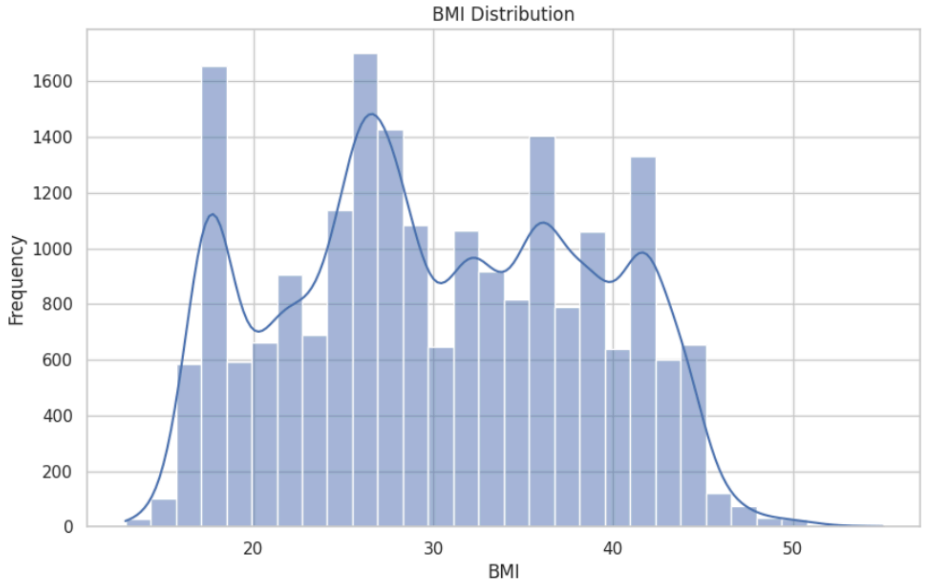


Fig. 7. BMI Distribution



Methods

Observation in our Study

Our project investigates how certain lifestyle factors can contribute to an individual's risk of developing obesity. Each observation in the data set represents an individual and their lifestyle characteristics which includes factors like dietary behaviors, such as calorie intake or the number of meals per day, and physical activity attributes, such as frequency of exercise and mode of transportation used. These attributes, alongside other relevant information such as water intake and alcohol intake, allow us to analyze the relationship between an individual's lifestyle factors and their risk of developing obesity.

Supervised vs. Unsupervised, Classification vs. Regression

Because our data set was generated from a deep learning model trained on an Obesity or CVD risk dataset, it is evident that there is a specific target variable, obesity level. Thus, we can infer that classification should be done since the target variable is being categorized into specific classes (e.g. "Normal_Weight", "Obesity_Type_I", "Overweight_Level_I").

Models and Algorithms

For this analysis, logistic regression can be employed for its interpretability and ability to identify the significance of various lifestyles that are attributed to obesity. It can help us to identify which lifestyle factors (like calorie intake or physical activity frequency) are statistically significant in predicting different obesity categories. Another mode of analysis can be decision trees which are ideal for capturing non-linear interactions, especially among dietary, physical, and lifestyle factors. These decision trees can reveal, for example, that individuals who exercise infrequently and consume a higher amount of calories are at a higher obesity risk. This will help

in understanding how multiple variables impact obesity. A kNN model can be used as a comparison model that can reveal that individuals with similar lifestyles may have a similar obesity risk.

Anticipated Weaknesses

One weakness that we anticipate being an issue is the distribution in the population size, known as a class imbalance. This could be troublesome if we decided to do a supervised learning model, leading to biased or inaccurate decisions amongst our clusters. If one population is overrepresented, the model will learn to prioritize it, leading to high accuracy for the majority class, but low accuracy on the minority. One way of combating this issue would be to undersample the majority class or oversample the minority to balance out the differences. An algorithm like a balanced random forest can apply higher weights made by errors in the minority class.

Feature Engineering

To prepare the data for our analysis, we had to encode several of the variables since the majority of them were categorical in nature. The target variable, NObeyesdad, had to be encoded to be on a scale of 0 to 6, with 6 being the most severe (Obesity_Type_III). The MTRANS variable had to be one-hot encoded, while the other categorical variables that had binary values could simply be mapped to a 0 or 1.

Results

Prediction Question

Our project revolved around the following prediction question: What factors predict obesity risk levels, and how do dietary and physical lifestyle attributes contribute to classification into the specific categories of obesity risk? This question highlights our goal in identifying key predictors of obesity risk to inform public health initiatives. For instance, health organizations can use these insights to design targeted interventions for populations similar to those represented in our data set. With the help of a model that can predict whether individuals are at risk of developing obesity, healthcare professionals can provide the necessary resources to help these individuals work towards reducing their risk of developing obesity which can potentially reduce the amount of future resources needed because of early intervention.

Linear Regression

Because the coefficients corresponding to each predictor variable in a linear regression model are more intuitive to understand, we thought that running a linear regression on our data would be a beneficial first step to determine whether we should continue with the linear model or explore other approaches. Although running the linear regression was not a part of the original pre-analysis plan, we thought it was worth investigating to further understand the nature of the data.

The linear regression model used an 80/20 split, and after 3 iterations of including a different number of predictors that we felt were most relevant by filtering out predictors that had coefficient values close to 0, we arrived at a model that includes the following predictors: family_history_with_overweight, FAVC, Weight, SMOKE, FAF, Gender, and Height (Fig. 8).

The R^2 of the model was 0.912, and the RMSE of the model was 0.616. In other words, the model explains about 91.2% of the variance in the target variable since values closer to 1 indicate a better fit, and the average error between the predicted and actual values is 0.59. Because the scale of the target variable ranges from 0 to 6, the predictions are roughly 0.59 units away from the true value within the context of the target variable's scale.

Fig. 8. Regression coefficients for linear regression model

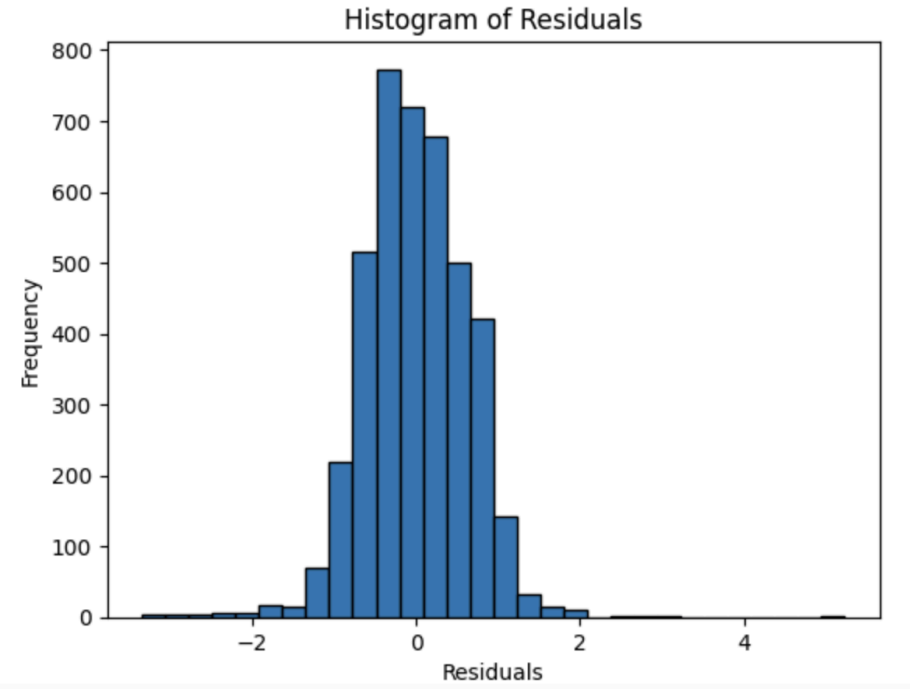
	variable	coefficient
0	family_history_with_overweight	0.362359
1	FAVC	0.070703
2	Weight	0.073685
3	SMOKE	-0.187604
4	FAF	-0.218952
5	Gender	0.463907
6	Height	-2.117179

For the positive coefficients, family_history_with_overweight and Gender seem to have a stronger influence on the target variable. For Gender, the values of "Female" were encoded to take on a value of 1 while the values of "Male" were encoded to take on a value of 0, and for family_history_with_overweight, the values of "no" were encoded to take on a value of 0 while the values of "yes" were encoded to take on a value of 1. Thus, we can see that the model predicts that an individual with a family history of being overweight will have a 0.36 higher target variable value and that a female individual will be predicted to have a 0.46 higher target variable value.

For the negative coefficients, height and FAF (physical activity frequency) seem to have a stronger influence on the target variable. As a person's height increases, the predicted value of the target variable decreases if all other factors are held constant. As an individual's physical activity frequency increases, the predicted value of the target variable also decreases.

To further look at the fit of the model, a histogram of the residuals can be analyzed to determine whether or not the model's errors are random. The histogram of residuals for the linear regression model appears to be normally distributed since it takes on a bell-shaped curve that is centered at 0 (Fig. 9). While there are some outliers at the far ends of the curve, they do not appear to be significant, so we can come to the conclusion that the model fits the data well.

Fig. 9. Histogram of Residuals



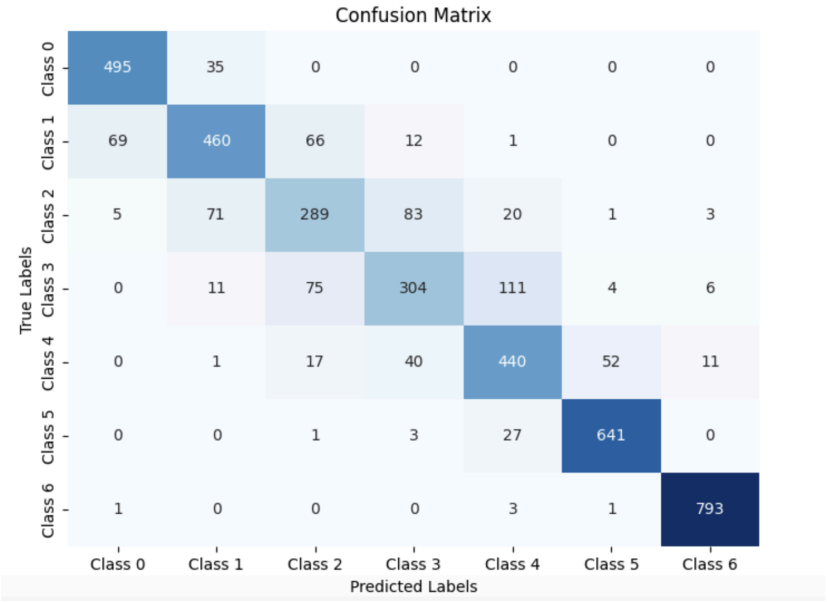
However, when looking at the range of the ŷ values, the minimum ŷ was -1.16, and the maximum ŷ was 9.26 which falls outside of the range of 0 to 6 of the target variable, indicating

that the model can still be improved. Thus, we decided to run a logistic regression which might be better apt at predicting categorical variables which is what our target variable is.

Logistic Regression

The logistic regression model included the same predictors as our linear regression model. Overall, the logistic regression model was reported to have an accuracy value of 0.824 which indicates that 82.4% of the samples were correctly classified. When looking at the confusion matrix, it is evident that class 6 had the most samples predicted correctly by the model (Fig. 10).

Fig. 10. Confusion matrix



The classification report also shows that class 6 had the highest precision and recall among all the other classes, so the model is particularly strong for predicting class 6 samples (Fig. 11). The classification report also shows that the model has high precision and recall for classes 0 and 5. The limitation of this model is that it seems to have difficulty classifying

samples within classes 2 and 3 which are values of “Overweight_Level_I” and “Overweight_Level_II” in the data based on their lower values in both recall and precision.

Fig. 11. Classification report

```
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.93      0.90       530
           1       0.80      0.76      0.78       608
           2       0.65      0.61      0.63       472
           3       0.69      0.59      0.64       511
           4       0.73      0.78      0.76       561
           5       0.92      0.95      0.94       672
           6       0.98      0.99      0.98       798
```

Conclusion

Complications

When working on the analysis portion of the project, we encountered some complications relating to the data itself and the algorithms we used. For the data we obtained from Kaggle, the entire dataset included two separate CSV files, one for testing, and one for training any models. Despite this convenient pre-separation of the data for a train-test split, the data for testing did not include any values for the target variable, NObeyesdad, so we had to resort to using the train portion of the dataset and further split that for our training and testing instead. Fortunately, the train portion of the whole dataset included 20,758 observations, so we were not limited by a small pool of data to do our analysis on.

For the linear regression model, one of the iterations of running the model included the modes of transportation as predictors. Since the MTRANS categorical variable was not binary,

we had to one-hot encode it in the feature engineering step. When we analyzed the values of the coefficients associated with each of the modes of transportation, we noticed that they all had large, positive values that were significantly greater than any of the other positive coefficients. Since it is illogical for all modes of transportation to significantly contribute to the value of the target variable, an explanation for the unexpected result of including the modes of transportation in the model is the manner in which they were one-hot encoded. Since all modes of transportation are included in the manner in which they are encoded, there might be the presence of collinearity. Thus, in the future, it would be better to reevaluate the ways in which we encoded the MTRANS variable and have one mode serve as the reference which can be dropped.

Limitations

Our data is representative of individuals from Mexico, Peru, and Colombia, so cultural, socioeconomic, and environmental factors influencing obesity risk may differ significantly in other regions. For instance, dietary patterns, access to healthcare, and levels of physical activity can vary widely across different populations. This model may not generalize well to other regions without incorporating broader determinants of health.

The dataset lacked variables such as socioeconomic status, mental health indicators, and other social determinants of health that are known to influence obesity risk. While variables like calorie intake/monitoring and physical activity allow us to assess lifestyle behaviors, they might act as proxies for more complex underlying factors such as access to healthy food or stress levels. This limitation means that the model might be missing several other critical factors that drive obesity.

Another concern lies in the ethical implications of using the model for decision-making. If it was deployed in real-world applications, the absence of variables like socioeconomic status or race might lead to biased outcomes, as they often correlate with health disparities. Conversely, including these variables could raise concerns about perpetuating systemic biases. Our analysis refrains from incorporating these variables, but future iterations of our research could explore different methods to address these concerns.

Works Cited

AravindPCoder. “Obesity or CVD Risk (Classify/Regressor/Cluster).” *Kaggle*, 20 Nov. 2023, www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster.

Chatgpt, chatgpt.com/. Oct. 2024.

Khil, Jaewon et al. “Water intake and obesity: By amount, timing, and perceived temperature of drinking water.” *PloS one* vol. 19,4 e0301373. 25 Apr. 2024, doi:10.1371/journal.pone.0301373.

Traversy, Gregory, and Jean-Philippe Chaput. “Alcohol Consumption and Obesity: An Update.” *Current obesity reports* vol. 4,1 (2015): 122-30. doi:10.1007/s13679-014-0129-4.