1. State a clear prediction question, not just what you did. Asking clear questions is the key to getting clear answers. You don't need to pick a "client", but sometimes it's helpful to think about the hypothetical scenario in which, say, NOAA or FEMA or a hospital would use your model.

Our prediction question is the following: What factors predict obesity risk levels, and how do dietary and physical lifestyle attributes contribute to classification into specific categories? This question addresses the need for 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.

2. Organize your discussion around two or three key tables or plots or numbers. Don't write up everything you did: Focus on your main result, and then add exposition around it until the reader is able to appreciate what you did and why. This keeps your writing from becoming bloated and rambling: Work backwards from the result to the reader's initial conditions.

We would want to include quantitative and categorical data. For quantitative we could use data like Height, Weight or age distribution. While for categorical reasons we could look into family history of being overweight another parameter that could be interesting is gender as a predictive factor. We did mention transportation as another categorical variable so perhaps we could also look into that.

- "The key metrics, F1-Score and Support, highlight the effectiveness of our models in predicting obesity categories. The F1-Score balances Precision and Recall, ensuring reliable performance across categories. For instance, the Decision Tree achieved exceptional F1-Scores, exceeding 98% in all categories, demonstrating its ability to minimize false positives and negatives.
- The Support metric, which indicates the number of instances in each category, shows that even for underrepresented groups like "Underweight" (283 instances), the models maintained high accuracy. These results validate the models' reliability, with Decision Trees performing particularly well, making them suitable for accurate and actionable obesity risk classification."

Logistic Regression

Category	**Precision**	**Recall**	**F1-Score**	**Support**
Underweight	0.855	0.961	0.905	283
Normal	0.879	0.824	0.85	466
Overweight	0.897	0.884	0.891	658
Obese	0.975	0.977	0.976	1361

There were a few limitations with each of our models. The logistic regression model struggled to distinguish between cases such as "Normal" and "Overweight" as reflected by a slightly lower recall for the "Normal" category. This limitation stems from the simplicity of the linear decision boundaries in logistic regression, which are less effective for overlapping data distributions.

Decision Tree

Category	Precision	Recall	F1-Score	Support
Underweight	0.986	1.0	0.993	283
Normal	0.987	0.985	0.986	466
Overweight	0.994	0.98	0.987	658
Obese	0.995	0.999	0.997	1361

The decision tree model is sensitive to small changes in the data, so if there were slight changes in variables like the Height or Weight when testing the dataset, the decision tree might output a different prediction than expected because the tree is so deep. This would not translate well when applying the model to real-world applications because there is natural variability in measurements relating to body metrics, and we cannot expect collected data to be perfectly accurate. Furthermore, when handling continuous variables such as Weight, the model can be prone to overfitting because there can be an issue of an excess number of splits when the continuous variables are not categorized into bins that are associated with some range of that variable. Furthermore, the decision tree model can lead to overfitting when a large number of discrete categories are involved since the model will tend to continue to split the leaf nodes in the tree if there is no restriction on the maximum number of leaf nodes.

Random Forests

Category	Precision	Recall	F1-Score	Support
Underweight	0.941	0.951	0.946	283
Normal	0.925	0.923	0.924	466
Overweight	0.942	0.956	0.949	658
Obese	0.994	0.985	0.99	1361

Random Forest models have a couple limitations that can impact their effectiveness and interpretability. One major drawback is their "black-box" nature. Since Random Forest predictions are based on averaging the outputs of multiple decision trees, it can be difficult to understand the importance of individual features or the logic behind specific predictions. While metrics such as precision, recall, and F1-score provide valuable performance insights, they do not explain why certain categories perform better or worse. Another limitation is class imbalance, which can skew the model's performance. For instance, in a dataset where the "Obese" category has significantly more support (1361 instances) compared to a minority class like "Underweight" (283 instances), the model tends to perform disproportionately well on the majority class. This imbalance often inflates precision, recall, and F1-scores for the dominant category while reducing the model's ability to accurately predict the minority classes.

3. You can handle criticisms or concerns in the results section if they are relevant or interesting, but most probably belong in the conclusion.

Several challenges that we encountered were during the analysis portion of our data. Logistic regression was one of the primary models used due to its interoperability. However, its coefficients represent log-odds ratios, which can be challenging to interpret directly. For example, a coefficient indicating a positive log-odds variable means it increases the likelihood of being classified in a higher-risk category, but the actual magnitude of the relationship is less intuitive compared to linear regression. We addressed this by focusing on the direction (positive or negative) of the relationships and relative importance rather than exact numerical values.

Our data is representative of individuals from Mexico, Peru, and Colombia. 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.