CS 418- Rose Matcha

Predicting Obesity in Individuals

Brenda Leyva, Filip Toloczko, Kevin Jara, Andy Wang, Sherwin Tahernezhadi

bleyva3@uic.edu | ftolo2@uic.edu | kjara2@uic.edu | awang72@uic.edu | stahe3@uic.edu

github.com/brendismode | github.com/filip-toloczko | github.com/kevinjara130 | github.com/Andy-Wang72 | github.com/Sherwin1600

Link to our project: https://github.com/uic-ds-fall2024/class-project-rose-matcha

Problem Statement

Problem Statement: Obesity is one of the most serious health problems in the world currently. It harms people's overall health and puts stress on healthcare systems. Our goal is to create a model that predicts whether or not an individual has obesity based on a variety of factors across individuals in the United States and Latin America. This model will help classify new cases more accurately and provide better information about what specific factors are important in causing this condition.

Why should others care? Learning about what specific factors lead to obesity can help people make healthier choices and guide doctors to focus on prevention. This can improve healthcare outcomes, make treatments more effective, and ideally help reduce costs for everyone.

How did we choose this problem? We chose this problem because obesity is a health issue that affects many people, including some of our own friends and family. We wanted to come up with some helpful insights to shine light on the nature of obesity as a whole.

Data

- We plan to use datasets from two populations.
  - The first dataset is data published by the CDC on health indicators for diabetes and obesity in the United States
  - The second dataset contains data on obesity and factors related to it for individuals from the Latin American countries of Peru, Colombia, and Mexico
- We have access to both of these datasets in the form of a CSV file. Both datasets were accessed via the UC Irvine Machine Learning Repository.
  - As such, the data collection process was not significantly challenging
- The CDC dataset consists of 253,680 rows with 22 columns/features and the Latin America dataset consists of 2,111 rows with 16 columns/features.
- Majority of feature columns in the CDC dataset are classified as float64 datatype, but are specifically binary variables taking on values of either 0 or 1.
  - E.g the smoker column consists of two values 0 and 1. 0 indicates that an individual is not a smoker and 1 indicates that the individual is a smoker.

Additional Considerations:

- Both datasets as presently constructed do not have an indicator variable for obesity. As such, we will build the obesity indicator variable from BMI where BMI > 30 will be classified as obese and non-obese otherwise
- Our eventual goal is to predict obesity based on various feature variables using ML methods. Note that our obesity target variable in the CDC dataset has some non-negligible class imbalance:
  - Obesity indicator has ~65% of the samples as non-obese and ~35% as obese
- We should address these class imbalance issues to ensure our ML models are not biased to be more likely to predict instances of the majority class as that would have a negative effect on model accuracy in predicting new/unseen samples.

Solution - Overview

- Our solution involved creating Logistic Regression and Decision Tree Classifier models for both the CDC and Latin America datasets to predict whether or not an individual is obese.
- We used K fold cross validation with k = 10 to evaluate the performance of our models and ensure they generalize well to unseen data.
- From the Decision Tree Classifier, we can also obtain feature importance scores for each feature which will tell us how important each feature is in terms of predicting obesity in both datasets.
- We also created visualizations to find any relationships between the factors that may help in exploring our proposed hypotheses and seeing if they are correct.
- After drawing inferences based on our visualizations and models we can conclude if our solution worked for our problem

Solution - Visualization

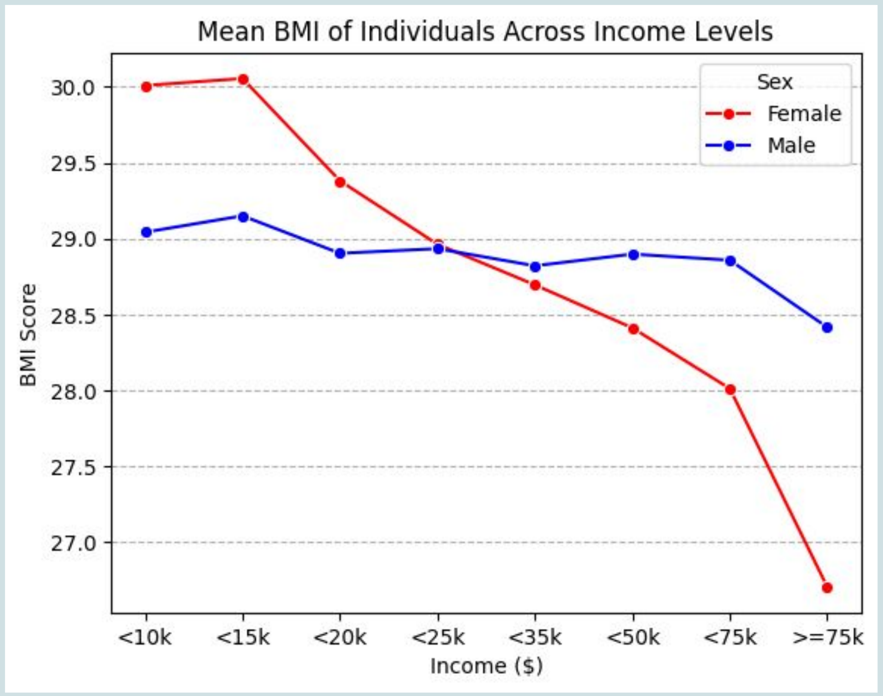
Our data allowed us to create some informative visualizations in order to test some interesting hypotheses. Some of the hypotheses we tested were:

- If an individual earns more money, then they will have a lower BMI.
- If an individual exercises, then they will have a lower BMI.
- If an individual eats healthy, then they will have a lower BMI.

These hypotheses allowed us to explore the impacts that several factors can have on the BMIs of individuals, which in turn tells us the obesity status of an individual. It is important to remember that in this context, anyone with a BMI greater than 30 is considered to be obese.

The following is a summary of each visualization, as well as a description of why we found each relationship visualized to be interesting along with what we learned from each hypothesis.

Solution - Visualization

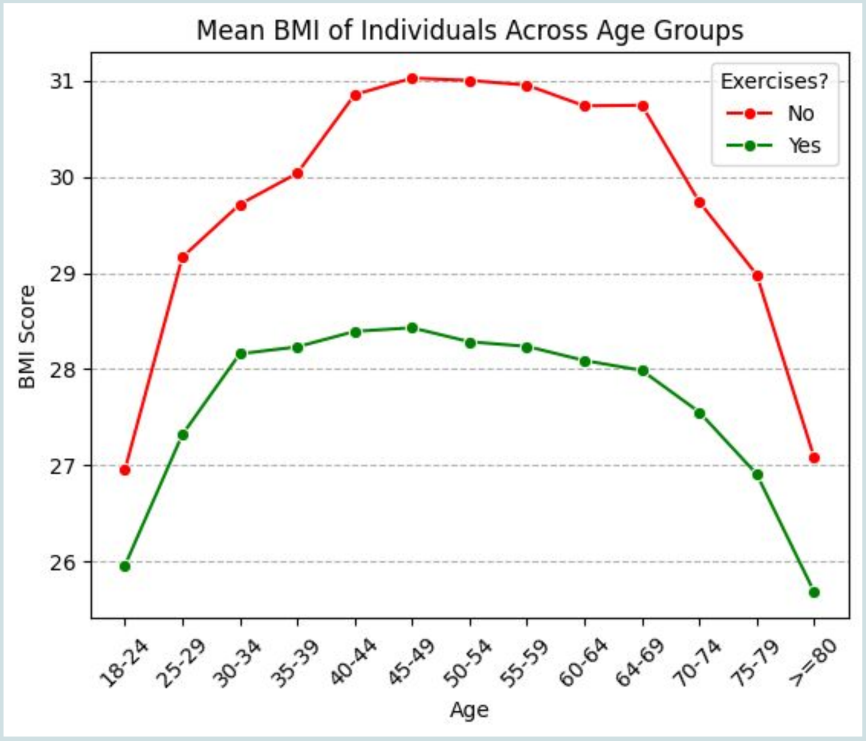


The first relationship that we would like to visualize, is the link between the income of an individual and their BMI. The hypothesis we had going into this was that a higher income would result in a lower BMI on average. This is because a higher income typically results in more access to healthy food as well as better medical access, likely reducing the chances of medical obesity.

This was in interesting hypothesis to test, as the income of an individual seemingly has a positive relationship with BMI. The higher an individual's income, the lower their BMI on average. While this is true across both sexes, it seemed to have a bigger impact on women. The impact on men is also apparent, with a clear downward trajectory, but with a much lower rate of change.

In summary the results of this graph show a relationship between income and obesity across both men and women in our test data. This implies that the lower your income, the closer the average person is to obesity, and at some incomes it even implies that the average person in that income bracket is obese.

Solution - Visualization

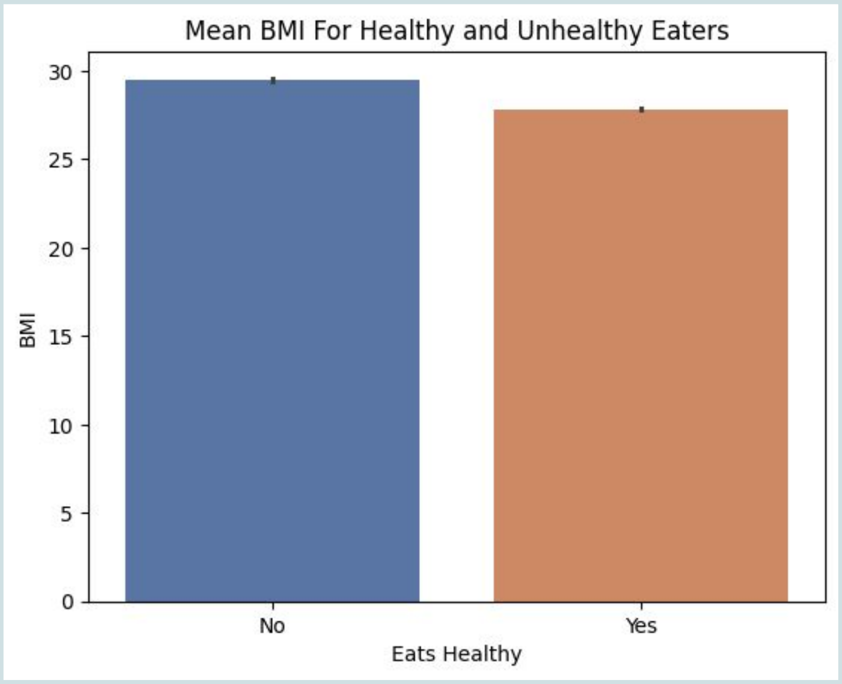


The second relationship that we would like to visualize is the link between exercise and BMI across age ranges. The hypothesis that we are going to be testing is that exercising is likely to lead to a decrease in an individual’s BMI. This is because exercising burns calories, and burning calories leads to losing weight. Therefore we expect that exercising lowers an individual's BMI and decreases the odds of them being obese.

This was an interesting hypothesis to test, as it not only confirmed our hypothesis, but also showed that it is true across all age ranges. We also found it interesting that the graph seems to show a similar curve across both groups, showing trends in BMI across age ranges regardless of exercise habits. It can also be said that regardless of exercise, young and old people had lower BMI scores that middle aged and upper middle aged people.

In summary, the results of this graph shows a relationship between obesity and exercise habits across age groups. For those who do not exercise and fall in certain age ranges, the average person in their group is obese. On the other hand, no age range in the exercising group has an average individual that is obese.

Solution - Visualization

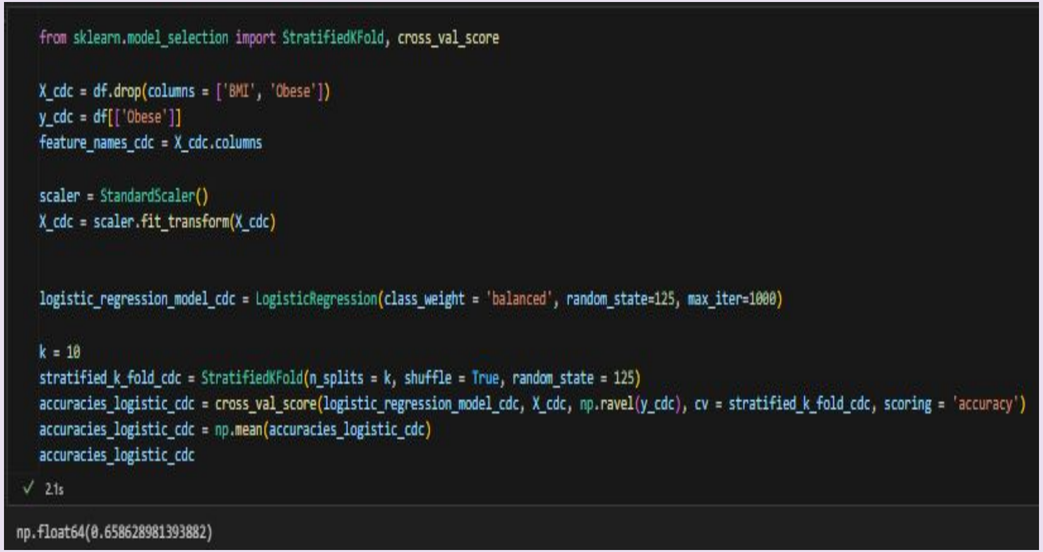


The third relationship that we would like to visualize is the link between eating healthy and BMI among individuals in our data set. The hypothesis that we are going to be testing is that eating healthy will lead to a lower BMI across all individuals. This is due to the fact that the variables we tested, fruits and veggies, are low in calories and tend to be full of healthy nutrients. We felt that those who did not have these groups present in their diets are more likely to be substituting them with less healthy, more calorie dense alternatives.

This was an interesting hypothesis to test, as the results of the visualization confirm our hypothesis, but not as much as expected. The difference in average BMIs was only a couple scores, less than we anticipated due to the fact that the mean of the group that doesn't eat healthy is under the obesity range.

In summary, the results of this graph shows a relationship between BMI and eating healthy. This implies that there are more obese individuals who do not eat healthy, but does not confirm that the average person with unhealthy eating habits is obese.

Solution - Machine Learning



```
from sklearn.model_selection import StratifiedKFold, cross_val_score

X_cdc = df.drop(columns = ['BMI', 'Obese'])
y_cdc = df[['Obese']]
feature_names_cdc = X_cdc.columns

scaler = StandardScaler()
X_cdc = scaler.fit_transform(X_cdc)


logistic_regression_model_cdc = LogisticRegression(class_weight = 'balanced', random_state=125, max_iter=1000)

k = 10
stratified_k_fold_cdc = StratifiedKFold(n_splits = k, shuffle = True, random_state = 125)
accuracies_logistic_cdc = cross_val_score(logistic_regression_model_cdc, X_cdc, np.ravel(y_cdc), cv = stratified_k_fold_cdc, scoring = 'accuracy')
accuracies_logistic_cdc = np.mean(accuracies_logistic_cdc)
accuracies_logistic_cdc
```

```
✓ 2.1s
```

```
np.float64(0.658628981393882)
```

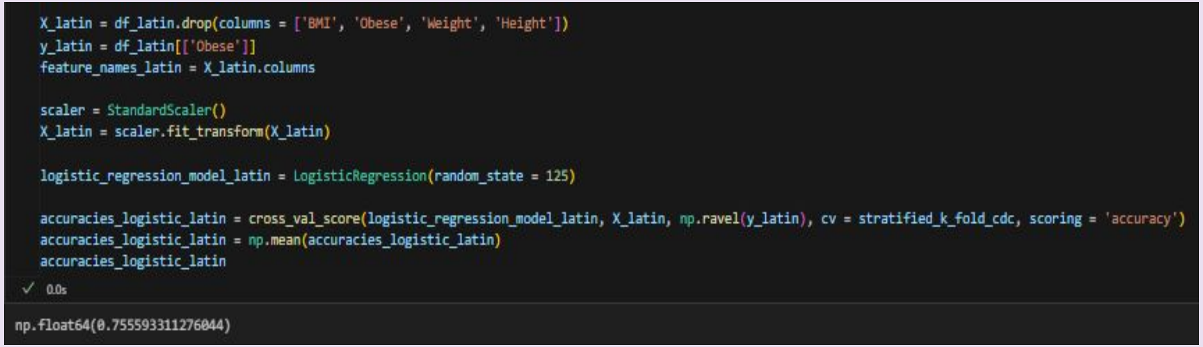
We will first run a model on main dataset, from the CDC. Due to having a binary response variable our first model for predicting obesity will be logistic regression since that classifier is well suited for binary classification and is simple and interpretable.

When defining the logistic regression model, we set the class weight parameter to ‘balanced’ to handle class imbalance issues in the target feature.

We use K-fold cross validation with k = 10 to evaluate our model on many different folds of the data and obtain an accuracy metric.

As can be seen from the output on the left, the logistic regression model had an accuracy of roughly .6586 in predicting the obesity of an individual from the CDC dataset (US Population)

Solution - Machine Learning



```
X_latin = df_latin.drop(columns = ['BMI', 'Obese', 'Weight', 'Height'])
y_latin = df_latin[['Obese']]
feature_names_latin = X_latin.columns

scaler = StandardScaler()
X_latin = scaler.fit_transform(X_latin)

logistic_regression_model_latin = LogisticRegression(random_state = 125)

accuracies_logistic_latin = cross_val_score(logistic_regression_model_latin, X_latin, np.ravel(y_latin), cv = stratified_k_fold_cdc, scoring = 'accuracy')
accuracies_logistic_latin = np.mean(accuracies_logistic_latin)
accuracies_logistic_latin
```

```
✓ 0.0s
```

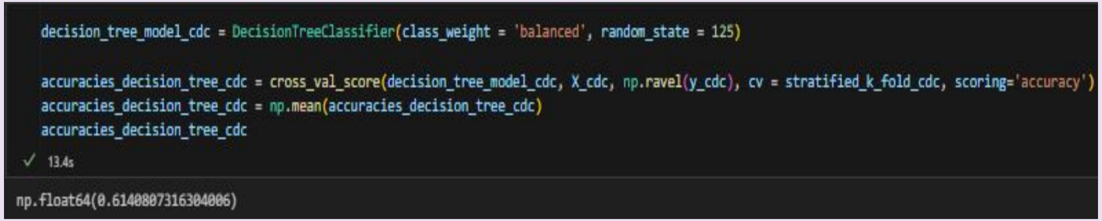
```
np.float64(0.755593311276044)
```

For our second dataset, Latin America, we similarly use K fold cross validation with k = 10 and obtain the resulting accuracy metric

After fitting a logistic regression model with Obese as the target, we got a accuracy rate of roughly 0.7555 (Latin America Population)

Solution - Machine Learning

CDC Decision Tree Classifier



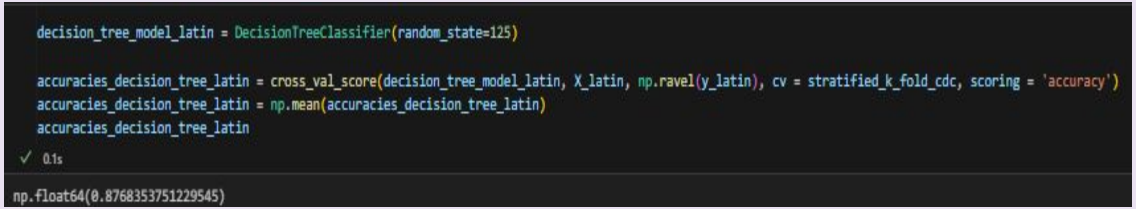
```
decision_tree_model_cdc = DecisionTreeClassifier(class_weight = 'balanced', random_state = 125)

accuracies_decision_tree_cdc = cross_val_score(decision_tree_model_cdc, X_cdc, np.ravel(y_cdc), cv = stratified_k_fold_cdc, scoring='accuracy')
accuracies_decision_tree_cdc = np.mean(accuracies_decision_tree_cdc)
accuracies_decision_tree_cdc
```

✓ 13.4s

```
np.float64(0.6140807316304006)
```

Latin America Decision Tree Classifier



```
decision_tree_model_latin = DecisionTreeClassifier(random_state=125)

accuracies_decision_tree_latin = cross_val_score(decision_tree_model_latin, X_latin, np.ravel(y_latin), cv = stratified_k_fold_cdc, scoring = 'accuracy')
accuracies_decision_tree_latin = np.mean(accuracies_decision_tree_latin)
accuracies_decision_tree_latin
```

✓ 0.1s

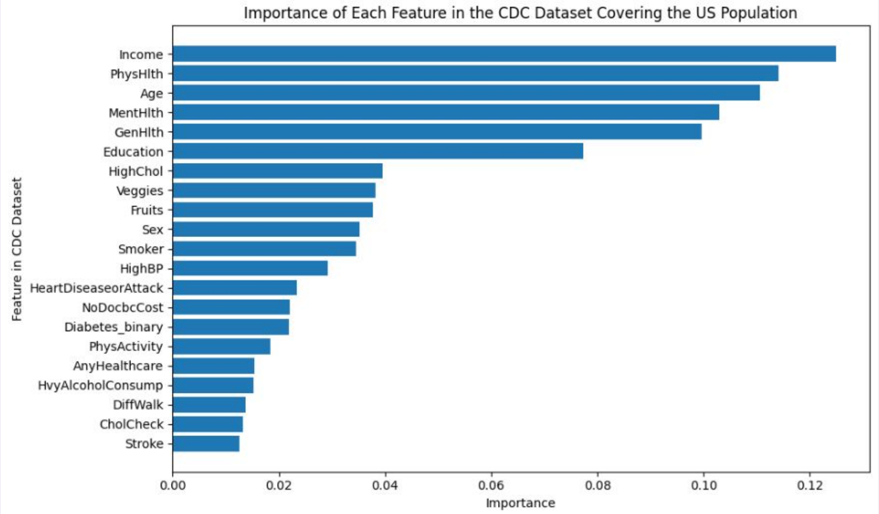
```
np.float64(0.8768353751229545)
```

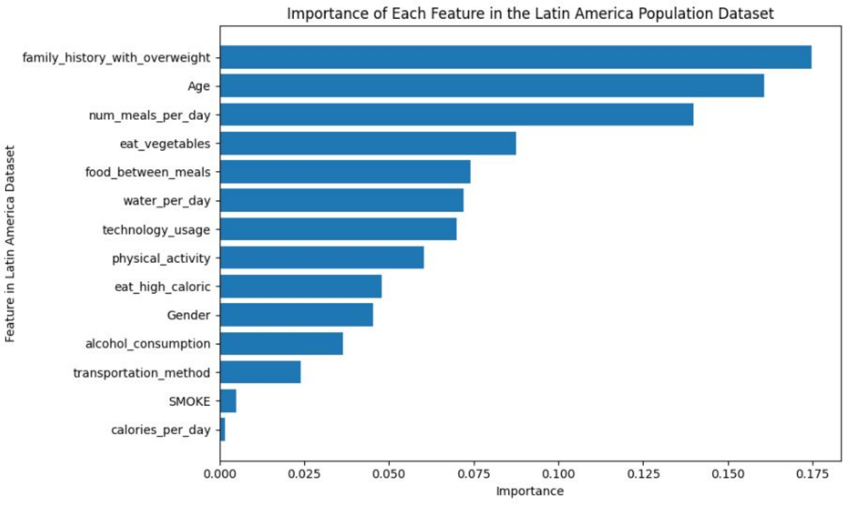
We also used the decision tree classifier to predict the obesity of an individual in both the CDC and Latin America datasets. Similarly to logistic regression, we utilized K fold cross validation and obtained our overall accuracy metric.

For the CDC dataset in predicting the obesity of an individual from the US population, the decision tree classifier had an accuracy of roughly 0.6140

For the Latin America dataset in predicting the obesity of an individual from the Latin America population (Mexico, Peru, Colombia), the decision tree classifier had an accuracy of roughly 0.8768

Solution - Machine Learning





Next we will attempt to find the features that are most significant in predicting obesity from both datasets. We did this by using sklearn's built in feature_importances_ method for the decision tree classifier and applying that method to our decision tree classifiers for both the CDC and Latin America datasets.

The most important features for each dataset are seen in the plots above, where the left plot represents the most important features for predicting obesity in the CDC dataset and the right plot represents the most important features for predicting obesity in the Latin America dataset.

As can be seen from the CDC plot, the most important features in predicting obesity in the US population are an individual's income, physical health, age, mental health, and education among others. In the Latin America plot, the most important features appear to be an individual's family history with being overweight, age, number of meals eaten per day, whether or not they eat vegetables, among others.

Between the two datasets, we can see some features that are common in being important in predicting obesity, like age and vegetable consumption. This finding would seem to indicate that regardless of underlying population (i.e. US or Latin American), an individual's age and vegetable consumption plays an important role in their obesity status.

Evaluation

Overall, it appears that our solution worked well. This is because we were able to create classification machine learning models to predict the obesity of an individual across two underlying populations (US and Latin America) with respectable accuracies hovering between roughly 60-87%, depending on the specific model and population. For instance, our logistic regression for the US population had an accuracy of roughly 65.86% while our decision tree classifier model for the Latin America population had an accuracy of roughly 87.68% which indicates that our models worked well in predicting obesity.

Furthermore, we also wanted to see which factors were most important in predicting obesity in each specific population. We successfully accomplished this through our use of feature importances from the decision tree classifier and subsequent construction of the feature importance plots seen on the previous slide.

All in all, our overall goal was to accurately predict the obesity of an individual from both the United States and Latin America populations in addition to understanding the most important factors in predicting obesity in both populations. As previously stated, we were able to accomplish this through our solution methodology thus our solution worked well overall.

Thank you for watching!

Thank you for watching and listening to our presentation, we hope that you found our findings interesting. We believe that this topic is one that is important, and that the issues we highlighted are something that more people should be aware of. We hope that the combination of machine learning and visualization techniques has been informative, and that you all learned something new.