Decision Tree Classifier and KNN Classifier for Overweight Classification

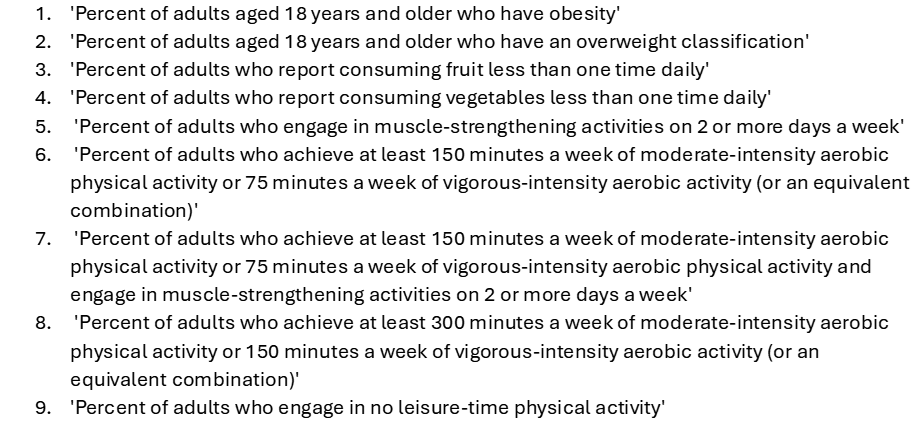
Our goal is to find how different attributes relate to the classification of a state population being on the higher or lower side of the percent of a population that is overweight.

Our Data was retrieved from Kaggle via this link: https://www.kaggle.com/datasets/spittman1248/cdc-data-nutrition-physical-activity-obesity

All available attributes with their non-null count and datatype are presented below.



All Records with a null ‘Data\_Value’ will be removed using the code: df = df.dropna(subset='Data\_Value').reset\_index(drop=True)

The ‘Questions’ column contains 9 unique questions that were asked of adults. The different questions can be seen below:

Each question has a corresponding value. To find the overall average value of each question for each state, I will filter the StratificationID1 to include only ‘Overall’; group the dataframe by state and question; aggregate the average ‘Data\_Value’ of each question for each state; then pivot our data frame so the questions show as columns without states being the row index. Python code for the transformations and a view of the head of the data frame/descriptive aggregate statistics can be seen below.

df = df[df['StratificationID1'] == 'OVERALL']

grouped\_df = df.groupby(['LocationDesc', 'Question'])['Data\_Value'].mean().reset\_index()

pivot\_df = grouped\_df.pivot\_table(index='LocationDesc', columns='Question', values='Data\_Value', aggfunc='mean')



We can see here that our distribution of overweight percentage between states has some left skew. I will thus use the median as our classification boundary:

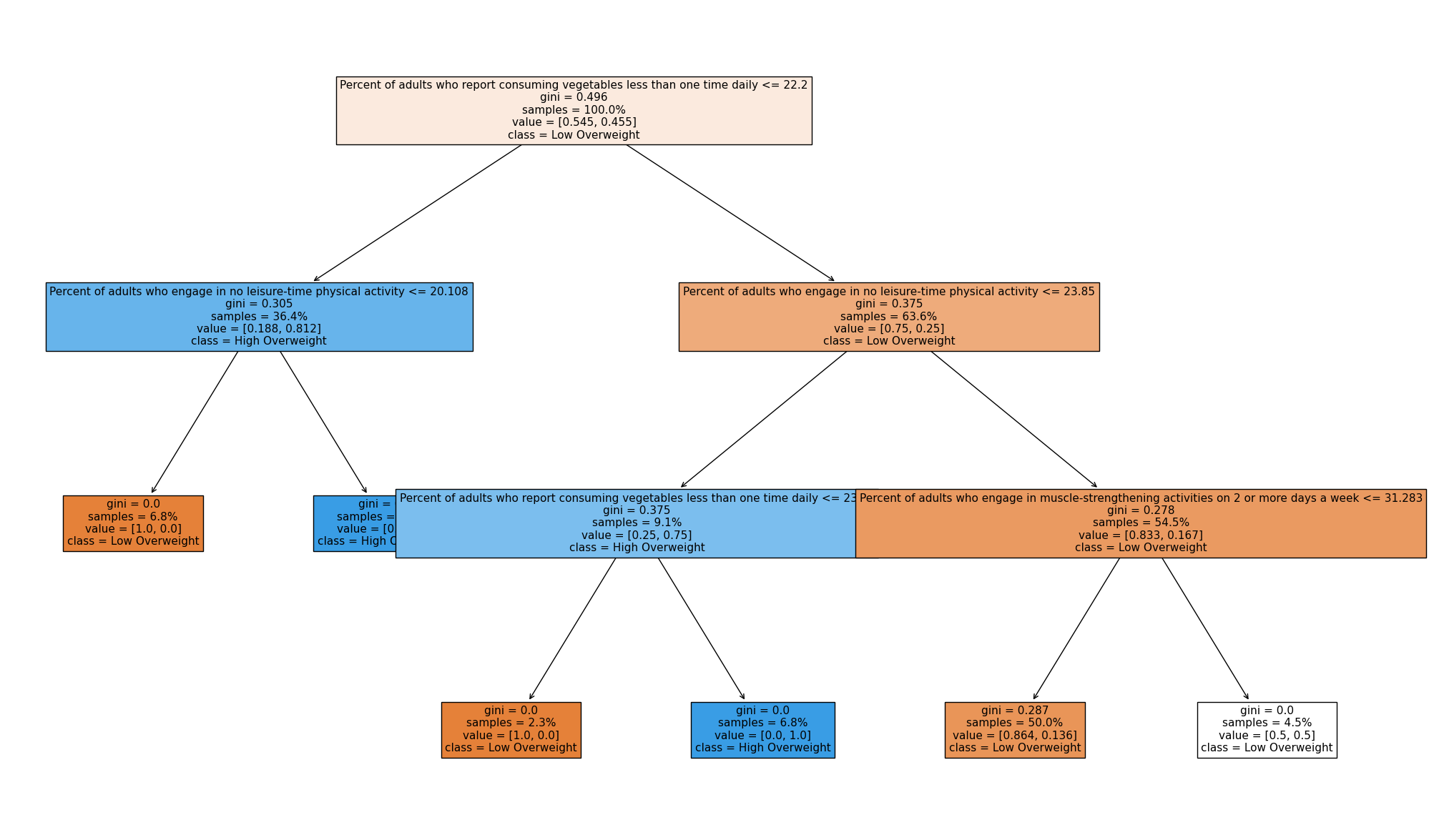
A graph of a number of adults

Description automatically generatedpivot\_df["High Overweight"] **=** pivot\_df['Percent of adults aged 18 years and older who have an overweight classification'] **>=** pivot\_df['Percent of adults aged 18 years and older who have an overweight classification'].median()

Code for our Decision Tree Classifier and the scores can be seen below.

A computer screen shot of colorful text

Description automatically generated

Our Decision Tree Visual Can be seen below

Our decision tree gives us a rather intuitive view to help determine if a state belongs in the high overweight class. Our primary indicator in this instance is the percentage of adults who report consuming vegetables less than one time daily---if true for more than 22.2 percent of the population, the model expects a high overweight class unless the percent of adults who engage in no leisure-time physical activity is greater than 20.1%. I wouldn’t expect more adults not participating in leisure-time physical activity to mean the class is more likely overweight. The sample size of 6.8% tells us this branch does not hold much generalizability.

The KNN Model and confusion Matrix can be seen below:

A screenshot of a computer

Description automatically generated

Note that there are two false positives and two false negatives. Initially, our Decision tree classifier seemed superior with better Accuracy, Precision, Recall, and F1 scores in the first model. The importance of cross-validation came into play as the mean scores across a 4-fold validation set show that the accuracy was the same while the Decision Tree had greater precision. However, cross-validation also showed the decision tree with worse recall and F1 score than expected. Both scores were quite comparable and variable, more model optimization would need to be done to find definite distinctions between model quality. The code for cross-validation and the table for scores and cross-validated scores are below.

