CS105 Final Project Report: Analyzing Obesity Statistics

By Sarah Ramirez, Jordan Sam, Laiba Hasan, Rohan Behera, Alex Szeto

Project Description/Proposal Link:
https://drive.google.com/file/d/1ECk91raX5myFmvosT1nMF_QPK12dy-XK/view?usp=share_link

1. Terms to know:

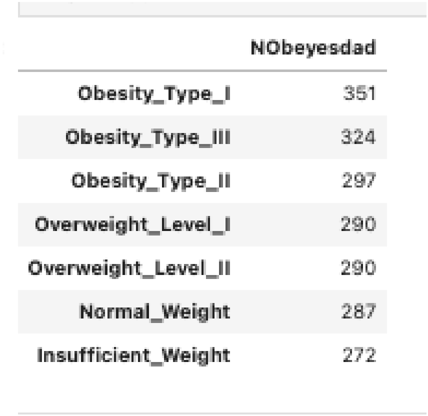
- FAVC- frequent consumption of high caloric food (yes or no)
- FCVC- frequency of consumption of vegetables (measured in 3 intervals: never, sometimes, always)
- NCP- number of main meals eaten per day (measured in 3 intervals: between 1 and 2, 3, more than 3)
- CAEC- consumption of food between meals (measured in 4 intervals: no, sometimes, frequently, always)
- SMOKE-if the participant smokes (yes or no)
- CH20- consumption of water daily (measured in 3 intervals: less than 1 liter, between 1 and 2L, more than 2L)
- CALC- consumption of alcohol (measured in 3 intervals: no, sometimes, frequently)
- SCC-calories consumption monitoring (yes or no)
- FAF-physical activity frequency (measured in 4 intervals: none, 1-2 days, 2-4 days, 4-5 days)
- TUE-time using technology devices (measured in 3 intervals: 0-2 hours, 3-5 hours, more than 5 hours)
- MTRANS-transportation used (automobile, motorbike, bike, public transportation, walking)
- NObeyesdad-obesity level (insufficient, normal weight, overweight level I, overweight level II, obesity type I, obesity type II, obesity type III)

2. Data collection/cleaning

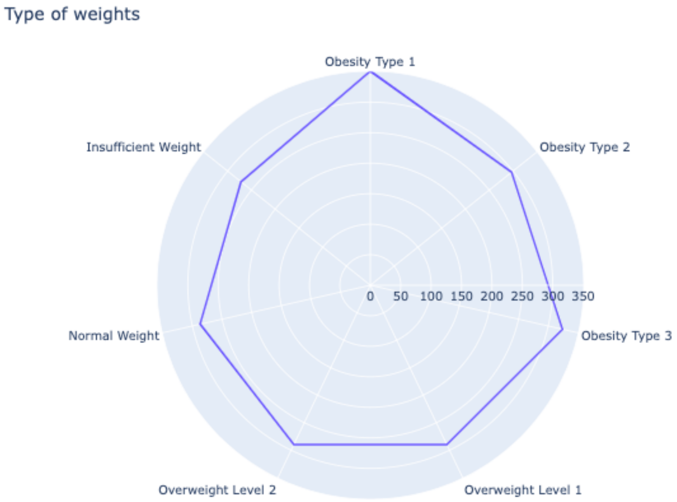
We rounded the age values to the nearest whole number, height and weight to the nearest two decimal places. CH20, FAF, and TUE were rounded to the nearest decimal place. We binned the ages into ranges of 2-5, 6-11, 12-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65 and older. We binned consumption of water per day as well into ranges of below normal, normal and above normal. We binned physical activity frequency as well into ranges of less than average, average, and above average. We binned the number of main meals a person consumes into two ranges: less than 3 meals and 3 or more meals. We created a new row calculating BMI from weight divided by height squared.

3. EDA

Visualization #1

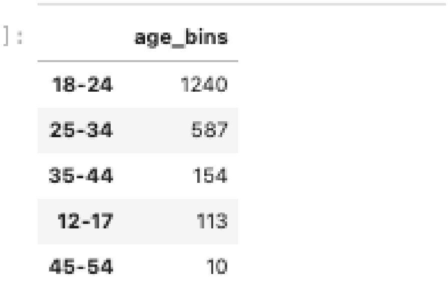


	NObeyesdad
Obesity_Type_I	351
Obesity_Type_III	324
Obesity_Type_II	297
Overweight_Level_I	290
Overweight_Level_II	290
Normal_Weight	287
Insufficient_Weight	272



Based on this spider chart we can see that the largest group of people fall under obesity type I. About 46% of all people surveyed fall under obese. About 27.47% of people fall under overweight.

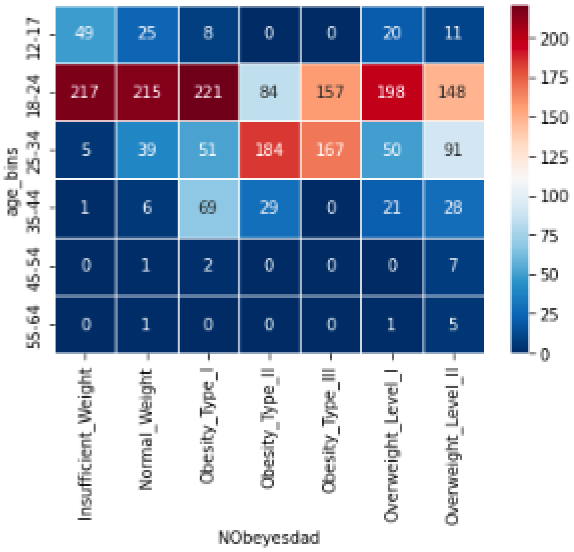
Visualization #2



]:

	age_bins
18-24	1240
25-34	587
35-44	154
12-17	113
45-54	10

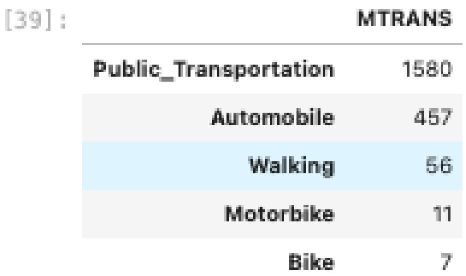
```
<AxesSubplot:xlabel='NObeyesdad', ylabel='age_bins'>
```



age_bins	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
12-17	49	25	8	0	0	20	11
18-24	217	215	221	84	157	198	148
25-34	5	39	51	184	167	50	91
35-44	1	6	69	29	0	21	28
45-54	0	1	2	0	0	0	7
55-64	0	1	0	0	0	1	5

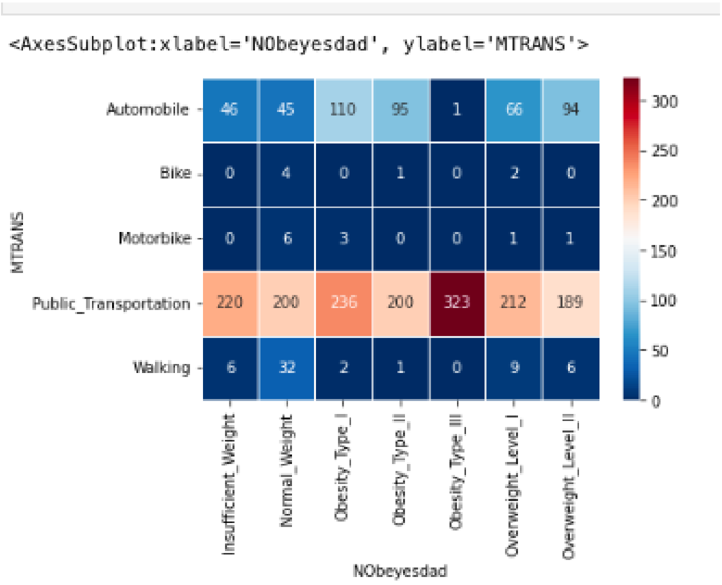
Based on this heat map we can see that the majority of people surveyed fall under the 18-24 age range with the exception of those in obesity type II and III in which case they fall under the age range of 25-34. This may be due to the fact that younger people are less likely to cook their meals and mostly dine outside at fast food restaurants and other dining establishments. They are also likely to consume more unhealthy snacks and beverages.

Visualization #3



[39]:

	MTRANS
Public_Transportation	1580
Automobile	457
Walking	56
Motorbike	11
Bike	7

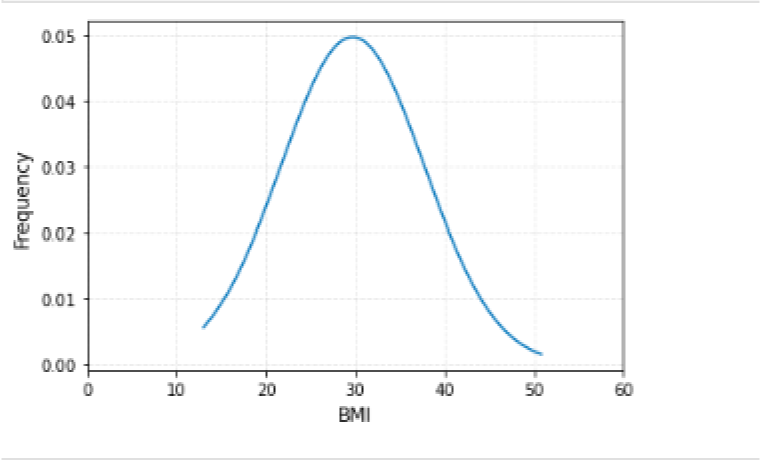


```
<AxesSubplot:xlabel='NObeyesdad', ylabel='MTRANS'>
```

MTRANS	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Automobile	46	45	110	95	1	66	94
Bike	0	4	0	1	0	2	0
Motorbike	0	6	3	0	0	1	1
Public_Transportation	220	200	236	200	323	212	189
Walking	6	32	2	1	0	9	6

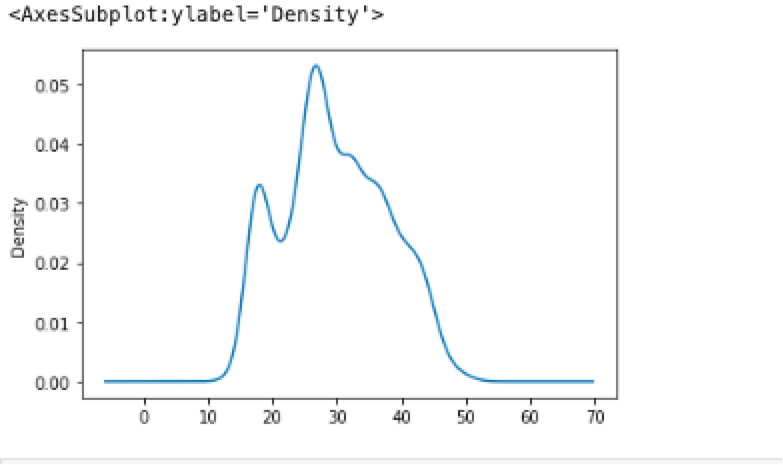
Although we see that there are a lot of people utilizing public transportation, it is important to note that we are looking at a dataset from Columbia, Peru, and Mexico where traveling is not as automobile. We also need to remember that this is 3 years ago which is pre COVID-19 pandemic times.

Visualization #4



Based on this graph, we can see that the BMI range of 25-35 has the highest frequency and the BMI of 45-50 has the lowest frequency.

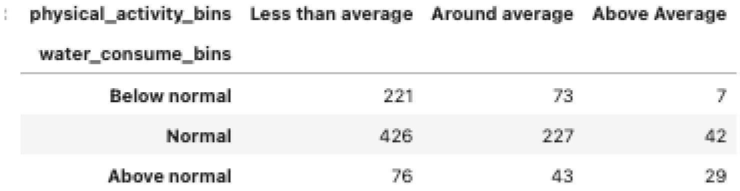
Visualization #5



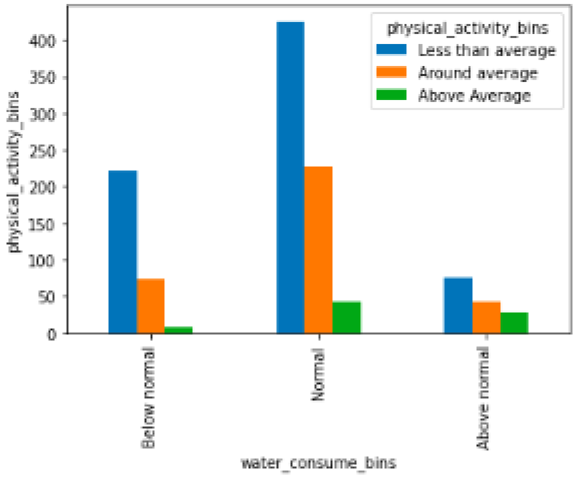
```
<AxesSubplot:ylabel='Density'>
```

Based on this graph we can see that the BMI density starts increasing at 12 and begins declining around 18 and goes up again at 20 and starts to decline again at 27.

Visualization #6

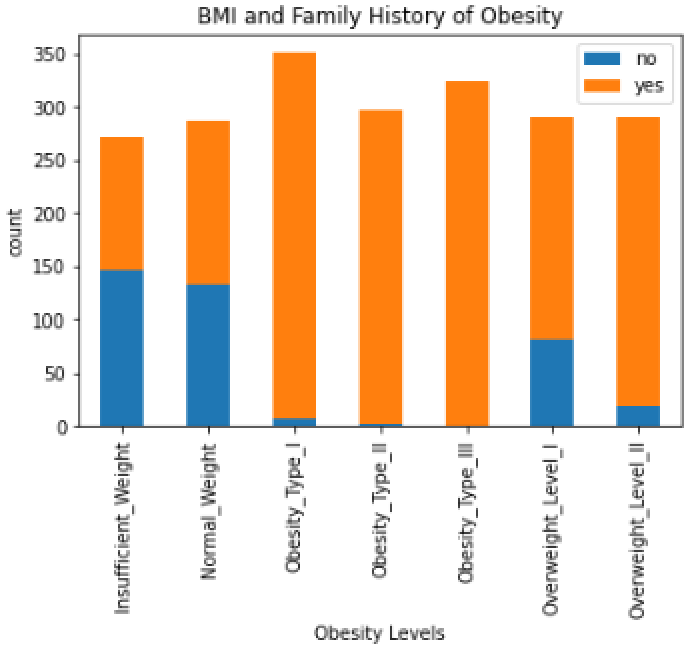


physical_activity_bins	Less than average	Around average	Above Average
water_consume_bins			
Below normal	221	73	7
Normal	426	227	42
Above normal	76	43	29



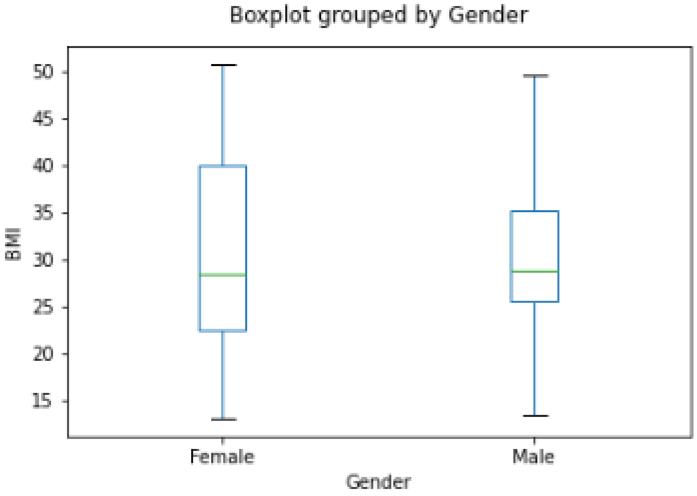
In this bar chart we can see that those who consume below normal amount of water daily, 221 people spend less the average amount of days outside exercising which accounts for roughly 71.42%% of people. For those who consume a normal amount of water daily, 426 people spend less than average amount of days outside exercising which accounts for 61.29%, followed by 227 people spending around average days exercising accounting for 32.66%, finally followed by 42 people spending above average exercising accounting for 6.04%. For people who drank above normal amounts of water, 76 of them spent less than average time outside, 43 spend around average, and 29 above average.

Visualization #7



This is a stacked bar chart where we have the counts of each obesity level and separate it by whether or not these people have a family history of obesity. We see that most of the families without a history of obesity are on the left of the graph in the insufficient weight and normal weight categories. This implies that people with a family history of being overweight have a higher likelihood of being obese.

Visualization #8

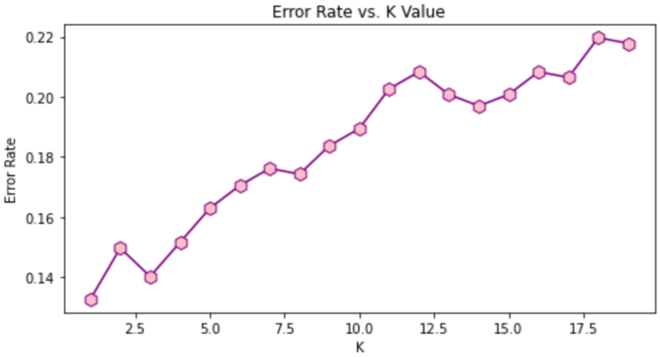


Looking at the boxplot we see that the mean BMI is about the same. The male's IQR is significantly smaller so the female BMI's have a much higher variability.

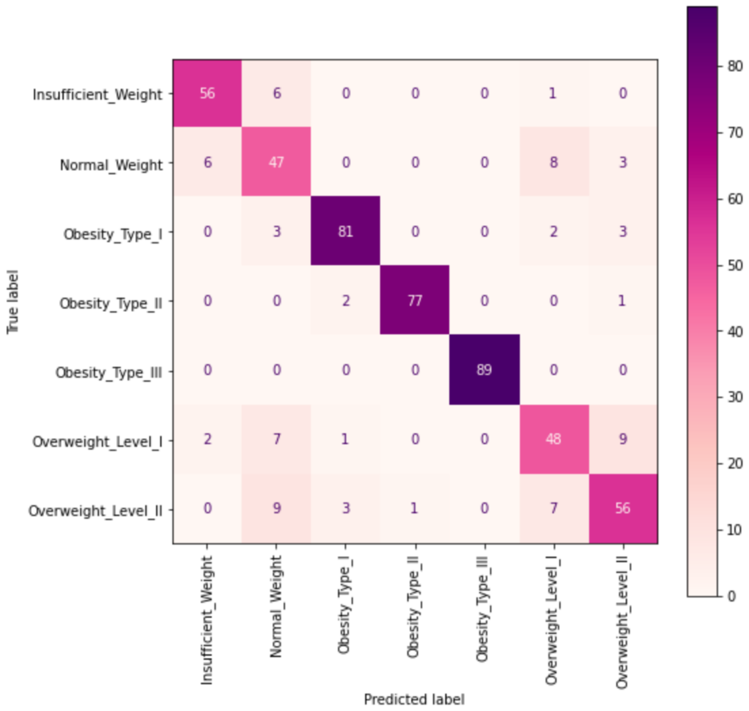
4. Main part

K-nearest-neighbors

We first started off by giving gender a numeric value. We gave Females the value “1” and Males the value “2”. We also assigned the frequent consumption of high caloric food (FAVC) responses with “1” for “yes” responses and “0” for “no” responses. From there we dropped columns that were unnecessary for classifying. These include the columns: NCP, CAEC, CH2O, CALC, SMOKE, SCC, TUE, MTRANS, BMI, age_bins , NObeyesdad, and family_history_with_overweight. After splitting the data into training and test sets we used the elbow method to find the K with the least error rate, in other words, the optimal K.



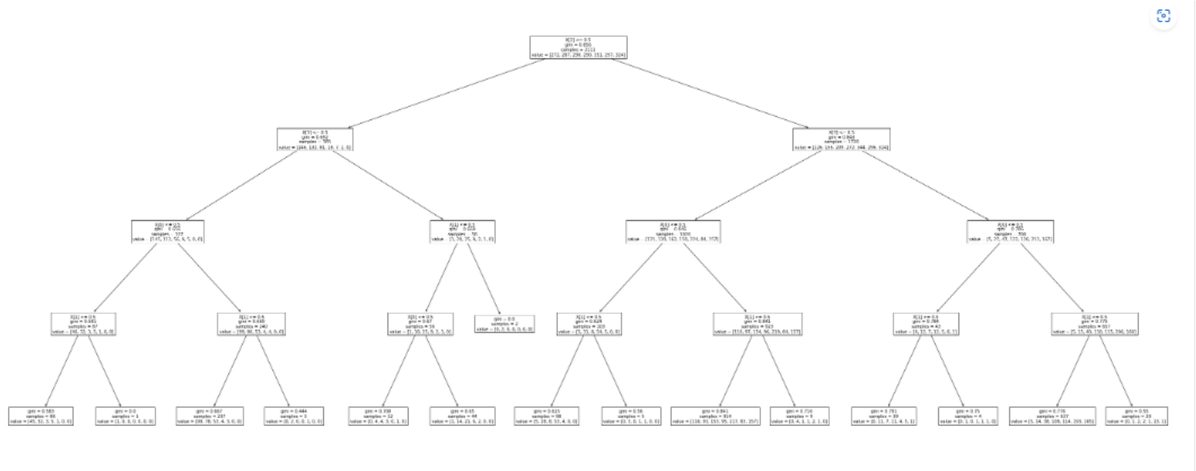
As you can see in the picture above the optimal K is 3. At K = 3, this is the last time the error rate drops before it continues to increase. After this we used Scikit Learn’s KNeighborsClassifier to find our KNN classifier using the optimal K we found before. Then we fit our KNN classifier from our training set. We then predicted the values for our X test set.



True label \ Predicted label	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Insufficient_Weight	56	6	0	0	0	1	0
Normal_Weight	6	47	0	0	0	8	3
Obesity_Type_I	0	3	81	0	0	2	3
Obesity_Type_II	0	0	2	77	0	0	1
Obesity_Type_III	0	0	0	0	89	0	0
Overweight_Level_I	2	7	1	0	0	48	9
Overweight_Level_II	0	9	3	1	0	7	56

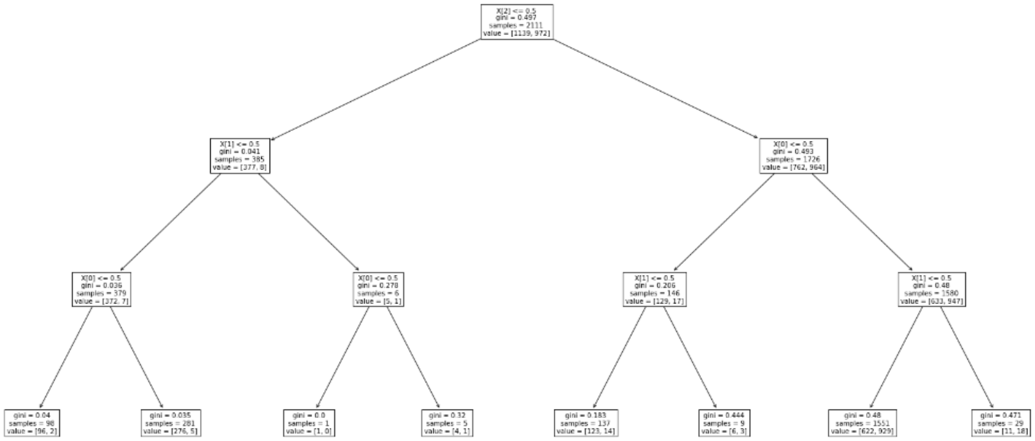
The diagonal line in the confusion matrix represents the true positive values for each category. In other words, it tells us how many obesity levels were predicted correctly from our test set.

Decision tree #1



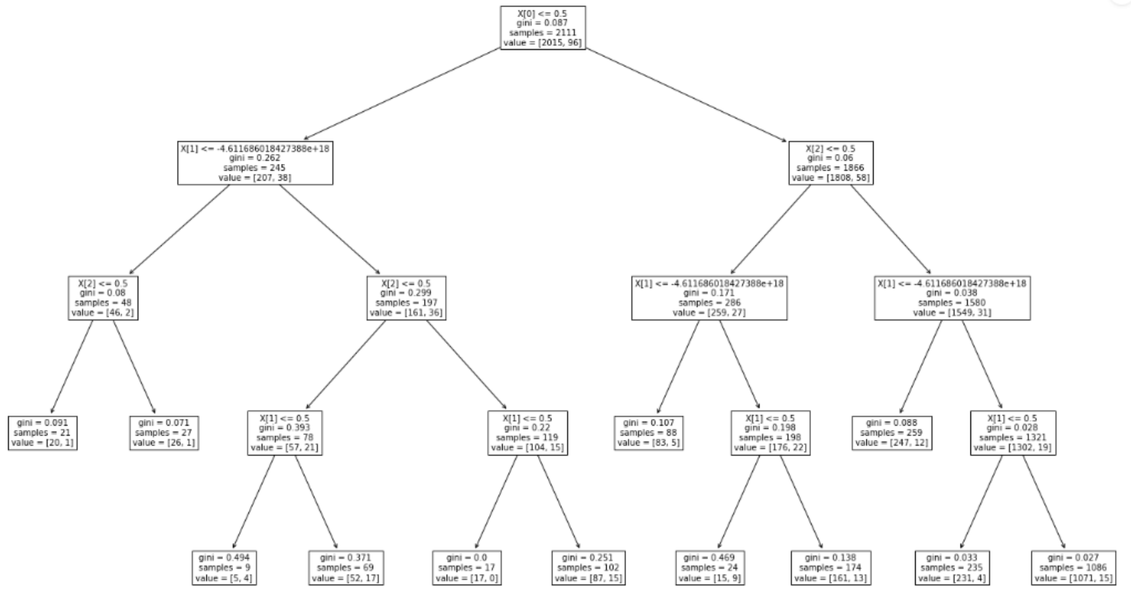
In this decision tree we have 4 features: smoking, high calorie food consumption, family overweight history, and age. If X[2] is less than or equal to 0.5 means if the user does not smoke, so the user will follow the true arrow to the left, if not they follow the false arrow on the right. Then we do the same thing for X[3] except this time the user does not consume high caloric food. For X[0] we do the same thing a third time except this time the user does not have family overweight history. For X[1] we do the same thing a fourth time except this time the user's age is under 25. Gini refers to the quality of the split, the samples represent the amount of people left in this decision. At the end of our decision tree we have 6 values. The first value in the array represents the amount of people from the sample with insufficient weight, the second value represents the number of people with normal weight, the third value represents the number of people with overweight_level_1, the fourth value represents the number of people with overweight_level_2, the fifth value represents the number of people with obesity_type_I, the sixth value represents the number of people with obesity_type_II, the seventh value represents the number of people with obesity_type_III.

Decision Tree #2



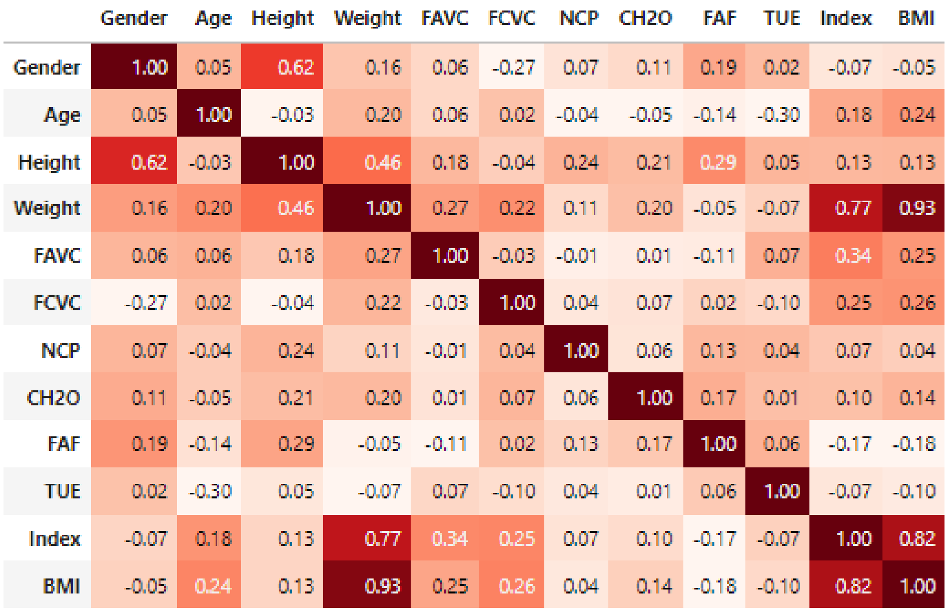
This decision tree is similar to the one above except we use 3 features this time: smoking, high calorie food consumption, and family overweight history to predict the obesity level. Also we binned insufficient weight, normal weight, overweight I, and overweight II into one range and the three obesity types into another range. So at the end of our decision tree we'll have two values: the one on the left represents the people that are not obese and the one on the right represents those that are.

Decision Tree #3



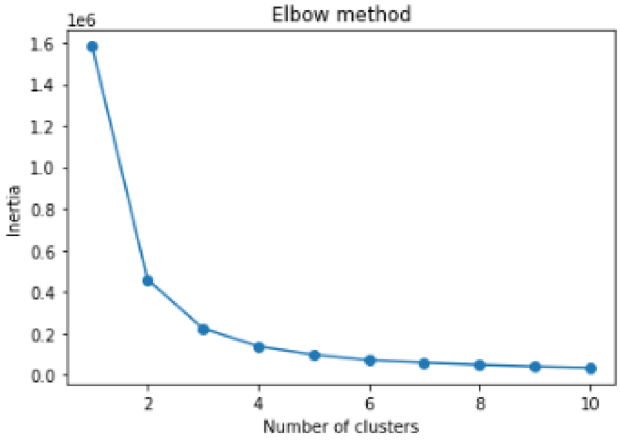
This decision tree is similar to the second one since we use 3 features, however this time we use the number of main meals consumed per day, high calorie food consumption and family overweight history to predict whether or not the user actually monitors the amount of calories they consume. At the end of our decision tree we have two values, the one on the left gives the number of users who track their calorie intake and the one on the right gives the number of users who don't.

K-Means Clustering

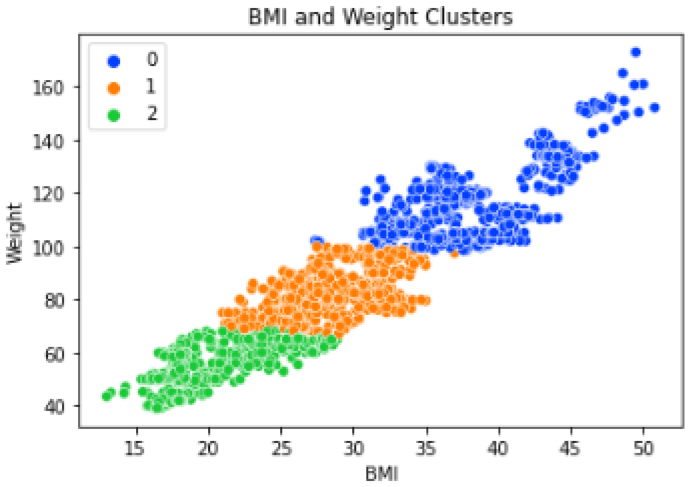


	Gender	Age	Height	Weight	FAVC	FCVC	NCP	CH2O	FAF	TUE	Index	BMI
Gender	1.00	0.05	0.62	0.16	0.06	-0.27	0.07	0.11	0.19	0.02	-0.07	-0.05
Age	0.05	1.00	-0.03	0.20	0.06	0.02	-0.04	-0.05	-0.14	-0.30	0.18	0.24
Height	0.62	-0.03	1.00	0.46	0.18	-0.04	0.24	0.21	0.29	0.05	0.13	0.13
Weight	0.16	0.20	0.46	1.00	0.27	0.22	0.11	0.20	-0.05	-0.07	0.77	0.93
FAVC	0.06	0.06	0.18	0.27	1.00	-0.03	-0.01	0.01	-0.11	0.07	0.34	0.25
FCVC	-0.27	0.02	-0.04	0.22	-0.03	1.00	0.04	0.07	0.02	-0.10	0.25	0.26
NCP	0.07	-0.04	0.24	0.11	-0.01	0.04	1.00	0.06	0.13	0.04	0.07	0.04
CH2O	0.11	-0.05	0.21	0.20	0.01	0.07	0.06	1.00	0.17	0.01	0.10	0.14
FAF	0.19	-0.14	0.29	-0.05	-0.11	0.02	0.13	0.17	1.00	0.06	-0.17	-0.18
TUE	0.02	-0.30	0.05	-0.07	0.07	-0.10	0.04	0.01	0.06	1.00	-0.07	-0.10
Index	-0.07	0.18	0.13	0.77	0.34	0.25	0.07	0.10	-0.17	-0.07	1.00	0.82
BMI	-0.05	0.24	0.13	0.93	0.25	0.26	0.04	0.14	-0.18	-0.10	0.82	1.00

A correlation matrix is a table that shows the correlation coefficient between all pairs of variables. This is an easy way to find patterns in the data. We look at the BMI row as we want to see what determines obesity status the most. The lighter spots show that frequent consumption of high caloric foods and frequent consumption of veggies have a .25 and .56 correlation with BMI. The darkest spot is weight with 0.93 correlation so we decided to use that look at that variable for clustering.



In k-means clustering we use the distance from the mean of the cluster to group data points. To determine the optimal number of clusters we tested k from 1:11. Having less clusters increases the inertia or intracluster distance while increasing k lowers it, with quickly diminishing returns. The elbow method is to use the number of clusters when the inertia flattens out, giving us k=3.



The algorithm starts by randomly assigning 3 data points to be the centroid and assigns surrounding data points to be in their cluster. The mean point of the clusters is then calculated and made the new centroid. This is repeated until the variance does not change, giving us 3 different groupings.

5. Contributions

What we all worked on:

- Picking the dataset for our project

- Notebook
- Report
- Slides
- Video
- Proposal
- Cleaning data: Discretized our data by binning (e.g numeric age becomes ranges (2-5, etc), Normalized our data, changed feature values to be numbers (e.g 'Yes' or 'No' to '1' or '0', or Gender to a arbitrary number)

Sarah Ramirez:

- Worked with Laiba on the KNN classification, we used training and test sets to train and test our model, used the elbow method to find the optimal K, created a confusion matrix and classification report for our model

Jordan Sam:

- Code / explanation for decision trees predictions (why the visualizations are meaningful)
- Worked with Rohan on decision trees logic, writing the code, and analyzing our results

Laiba Hasan:

- Worked with Sarah on the KNN classification, we used training and test sets to train and test our model, used the elbow method to find the optimal K, created a confusion matrix and classification report for our model

Rohan Behera:

- Worked with Jordan on using categorical variables to create three decision trees predicting if user is obese or not and if they track the number of calories they consume
- Worked on bar chart comparing amount of water users consumed and the time spent on physical activities

Alex Szeto:

- EDA and analysis, made the BMI stacked bar chart and box plot
- K-means clustering on BMI and weight (correlation matrix, elbow method, k-means clustering)

Presentation Video Link:

https://drive.google.com/file/d/1iCfre2fjAuWi_UTC6pjzLW7_aN0yH11U/view?usp=share_link