Analysis of Obesity Levels Using the Apriori Algorithm

Tim – Dec 9, 2020

INTRODUCTION

This report examines how a person's weight can be influenced by eating habits, physical condition and other factors. The dataset used for analysis is titled "Estimation of obesity levels based on eating habits and physical condition Data Set" and can be found at the following link. This dataset contains 2111 records with 17 attributes, however only 7 attributes will be considered here. The Apriori algorithm was used to analyse and break it into frequent itemsets in a Jupyter Notebook.

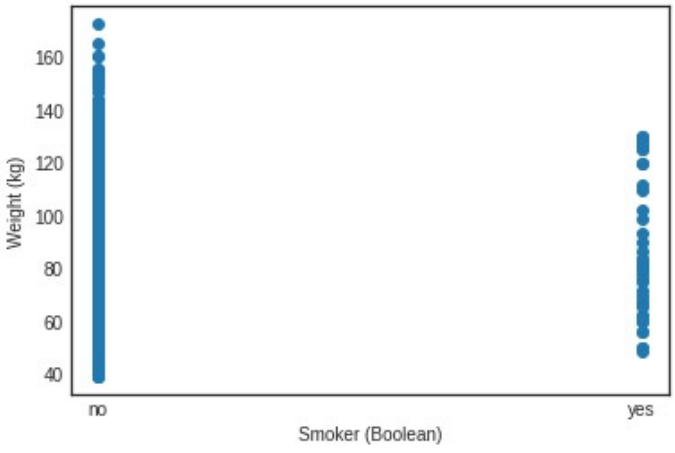
RESEARCH QUESTIONS

Using this dataset, I wanted to try and answer the following questions:

Are people that drive their car more likely to be overweight?
Are smokers more likely to be overweight?
Are people with a family history of obesity more likely to be overweight?
Are people who eat high caloric food more likely to be overweight?
Are people who keep track of their calories more likely to be overweight?
Are people who eat food between meals more likely to be overweight?

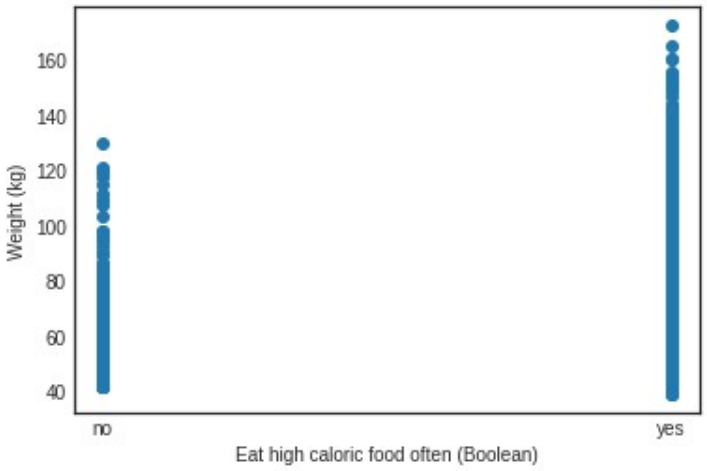
PRELIMINARY ANALYSIS

Before delving into the algorithm, I plotted the two variables against each other to try and guess the answers to my research questions.

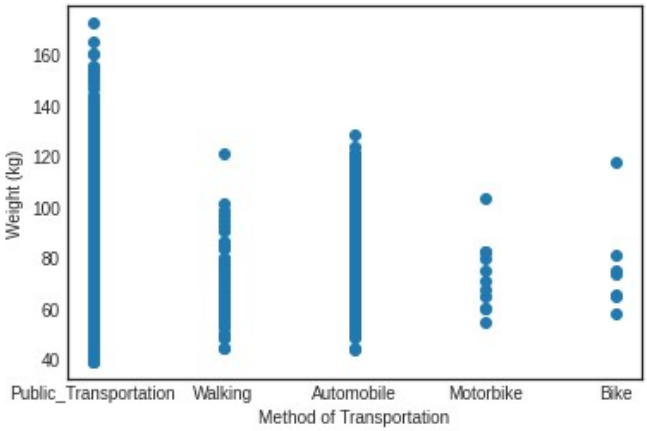


From the graph to the left, it appears that smokers are more likely to have a lower weight than non-smokers. Thus, smokers are less likely to be overweight.

This graph is evidence that people who have overweight people in their family history are more likely to be overweight.



People that consume high caloric food frequently are more likely to be overweight.



Not surprisingly, people that walk or bike to work are not as likely to be overweight. The people that are most likely to be overweight are people who use automobiles or public transit.

Keeping track of the calories one consumes can help people maintain a healthy weight as seen in the graph above.

APRIORI ALGORITHM

The following 7 predictors are selected from the dataset: family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, MTRANS and NObeyesdad (descriptions below). We will examine the frequent itemsets of this dataset using the Apriori algorithm to see if two behaviours or habits occur together. For example, if many overweight people eat high caloric food, we would expect to see this itemset more frequently. All the unique domain values for all 7 predictors is 21, so we have 2111 market baskets and 21 items.

I need to modify the domains of the attributes I am examining because I cannot compare itemsets with the same values, so I have to give unique values to each attribute. I shortened some of the value names in the modified domains to speed up the algorithm.

The Apriori algorithm considers only the larger sets whose subsets are frequent, so we can only select doubletons that are made up of frequent singletons. We will run the algorithm with a threshold of 200.

Attributes	Description	Domains	Modified Domains
Height		Floating point number in meters	
Weight		Floating point number in kilograms	
family_history_with_overweight	Family history of overweight	{‘Yes’, ‘No’}	{‘fh_y’, ‘fh_n’}
FAVC	Do you eat high caloric food frequently?	{‘Yes’, ‘No’}	{‘fa_y’, ‘fa_n’}
CAEC	Do you eat any food between meals?	{‘No’, ‘Sometimes’, ‘Frequently’, ‘Always’}	{‘cae_no’, ‘cae_so’, ‘cae_fr’, ‘cae_al’}
SMOKE	Do you smoke?	{‘Yes’, ‘No’}	{‘sm_y’, ‘sm_n’}
SCC	Do you monitor the calories you eat daily?	{‘Yes’, ‘No’}	{‘sc_y’, ‘sc_n’}
MTRANS	Which transportation do you usually use?	{‘Automobile’, ‘Motorbike’, ‘Bike’, ‘Public_Transportation’, ‘Walking’}	{‘au’, ‘mo’, ‘bi’, ‘pu’, ‘wa’}
NObeyesdad	Obesity level	{‘Insufficient_Weight’, ‘Normal_Weight’, ‘Overweight_Level_I’, ‘Overweight_Level_II’, ‘Obesity_Type_I’, ‘Obesity_Type_II’, ‘Obesity_Type_III’}	{‘ins’, ‘nrm’, ‘ovr’, ‘ob’}

The NObeyesdad variable is simply the body mass index (BMI). Body mass index = weight/(height*height).

PROBLEMS

After running the algorithm with a threshold of 200, I was not getting good results that included the variable NObeyesdad. So I restricted the obesity levels to insufficient, normal, overweight and obese. This gave me better results. Initially, I included gender as a predictor, but I did not find many interesting results, so I eliminated it. I had implemented hashing, however, this turned out to slow the algorithm down and many doubletons were being hashed to the same bucket, so I stopped using it for doubletons.

RESULTS

Are people that drive their car more likely to be overweight?
item:['ob', 'pu'] frequency:758
item:['ovr', 'pu'] frequency:400
normal and ins are 2111 – (758 + 400) = 953
Are smokers more likely to be overweight?
item:['sm_n', 'ob'] frequency:949
item:['sm_n', 'ovr'] frequency:571
normal and ins are 2111 – (949 + 571) = 591
Are people with a family history of obesity more likely to be overweight?
item:['fh_y', 'ob'] frequency:963
item:['fh_y', 'ovr'] frequency:480
normal and ins are 2111 – (963 + 480) = 668
Are people who eat high caloric food more likely to be overweight?
item:['ob', 'fa_y'] frequency:952
item:['ovr', 'fa_y'] frequency:483
item:['nrm', 'fa_y'] frequency:207
item:['ins', 'fa_y'] frequency:220
Are people who keep track of their calories more likely to be overweight?
item:['ob', 'sc_n'] frequency:968
item:['ovr', 'sc_n'] frequency:538
normal and ins are 2111 – (968 + 538) = 605
Are people who eat food between meals more likely to be overweight?
item:['ob', 'cae_so'] frequency:953
item:['ovr', 'cae_so'] frequency:505
normal and ins are 2111 – (953 + 505) = 653

From this data we can conclude that the answer to all of our research questions is yes, except the first, which needs more clarification. This matches almost all of the predictions we made from our preliminary analysis. We can expect people that take public transport to be more likely to be overweight. The number of obese people that drive an automobile was 205 compared to 758 for public transport.

Some other interesting results can be found in this data:

item: ['fh_y', 'sc_n']	frequency:1678	item: ['pu', 'fa_y']	frequency:1404
item: ['fh_y', 'cae_so']	frequency:1545	item: ['sc_n', 'fa_y']	frequency:1807
item: ['fh_y', 'fa_y']	frequency:1579	item: ['sc_n', 'cae_so']	frequency:1710
item: ['pu', 'sc_n']	frequency:1505	item: ['cae_so', 'fa_y']	frequency:1607

There were 1678 occurences of people with a family history of obesity who didn't keep track of their calories. People with a family history of obesity were also more likely to eat food between meals and eat high caloric food frequently. Riding public transportation is a behaviour that is also correlated with not keeping track of calories consumed and eating high caloric food with 1505 and 1404 occurrences, respectively. People that eat high caloric food were also more likely to not keep track of their calories consumed. The following two frequent triples are implied and can be confirmed by calculating all frequent triples:

family history of overweight, not keeping track of calories and eating between meals:

['fh_y', 'sc_n', 'cae_so]

item: ['fh_y', 'sc_n']	frequency:1678
item: ['fh_y', 'cae_so']	frequency:1545
item: ['sc_n', 'cae_so']	frequency:1710

public transportation, not keeping track of calories and eating high calorie food:

['pu', 'sc_n', 'fa_y']

item: ['pu', 'sc_n']	frequency:1505
item: ['pu', 'fa_y']	frequency:1404
item: ['sc_n', 'fa_y']	frequency:1807

FURTHER RESEARCH

Binary classification could be used on the outcome variables Normal weight vs. the other outcomes. Finding the most frequent triples (or n-tuples) could help show correlation between variables, such as finding whether a person who doesn't smoke and uses public transport is more likely to be overweight. It could also confirm our suggestion that using public transportation, not keeping track of calories and eating high caloric food could be a frequent triple. The algorithm can be run with different thresholds that could show more results not shown here. Other variables can be added in such as the frequency of the use of alcohol. Obesity levels could also be clumped differently, such as into insufficient, normal and overweight.

CONCLUSION

A habit of eating high caloric food can be an indicator of obesity. Excess calories are stored as fat on the body, so people that consume more calories, are more likely to gain weight. People that bike to work are less likely to be overweight, however the sample size for bikers is small since not many people bike to work. Using an automobile or public transport is an indicator that a person might be overweight. This could be linked to the stress involved in waiting in traffic, getting to the bus or subway on time and the social anxiety of being near strangers in a small space. Stress and weight gain are known to be linked together. Smoking suppresses a person's appetite, but this is not a recommended way to lose weight. Lastly, people that eat food between meals and eat high calorie food are more likely to be overweight.