CS105 Final Project Proposal

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

Introduction: Our dataset contains 2111 columns and uses 16 different factors to estimate the level of obesity in a human. There are 7 different categories of obesity: underweight/insufficient, normal, overweight level I, overweight level II, obesity level I, obesity level II, obesity level III. 23% of the responses were taken from humans and the remaining 77% was generated with the Smote filter and Weka tool. We chose this dataset because we felt it was interesting to see how factors that cause obesity can differ in countries outside of the United States such as people's form of transportation.

Description: Our project analyzes certain factors such as physical activity, number of meals consumed per day, method of transportation, consumption of water, etc. and how they can influence whether or not a person has obesity. The data is taken from 3 latin american countries: Columbia, Peru and Mexico.

Dataset:
https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+

Visualizations we plan on making:

- Spider chart analyzing at the weight/obesity level the users surveyed fall under
- Heat map visualizing the correlation between weight/obesity level and the age range the users surveyed fall under
- Heat map visualizing the correlation between method of transportation and weight/obesity level
- Line graph showing the correlation between BMI and frequency
- Line graph showing the correlation between BMI and density
- Box plot analyzing the correlation between gender and BMI

Machine learning algorithms we plan on using:
The technique we plan on using for our project is classification. The supervised model we want to use is KNN in order to figure out if we can determine which level of obesity a person falls under based on their answers, in particular their gender and whether or not they consume high caloric food frequently. We plan to use a confusion matrix as well to analyze the true vs predicted values of obesity levels given k. We want to use feature selection as well to determine which categories will affect obesity levels the most as well as train the supervised model to make a decision tree. Features we will be using are smoking, high calorie food consumption, family overweight history, and age range. We think decision trees could work because some of our categories accept only yes or no for an answer, which we can easily convert to true or false. Other categories like age can be binned to a range of either under or over 25 years old to make them compatible with the decision tree. We also plan on using k-means clustering for BMI and weight.