CSE574 Milestone 1 Report

## Team Member

Name ： PinLyu Huangwei Ding Matthew Sah

Email ID: [pinlyu@buffalo.edu](mailto:pinlyu@buffalo.edu) [hding3@buffalo.edu](mailto:hding3@buffalo.edu) [msah@buffalo.edu](mailto:msah@buffalo.edu)

UB Person ID: 50321560 50155734 50320376

## Topic : Prediction of Diabetes and the Relevance to Dietary Habits

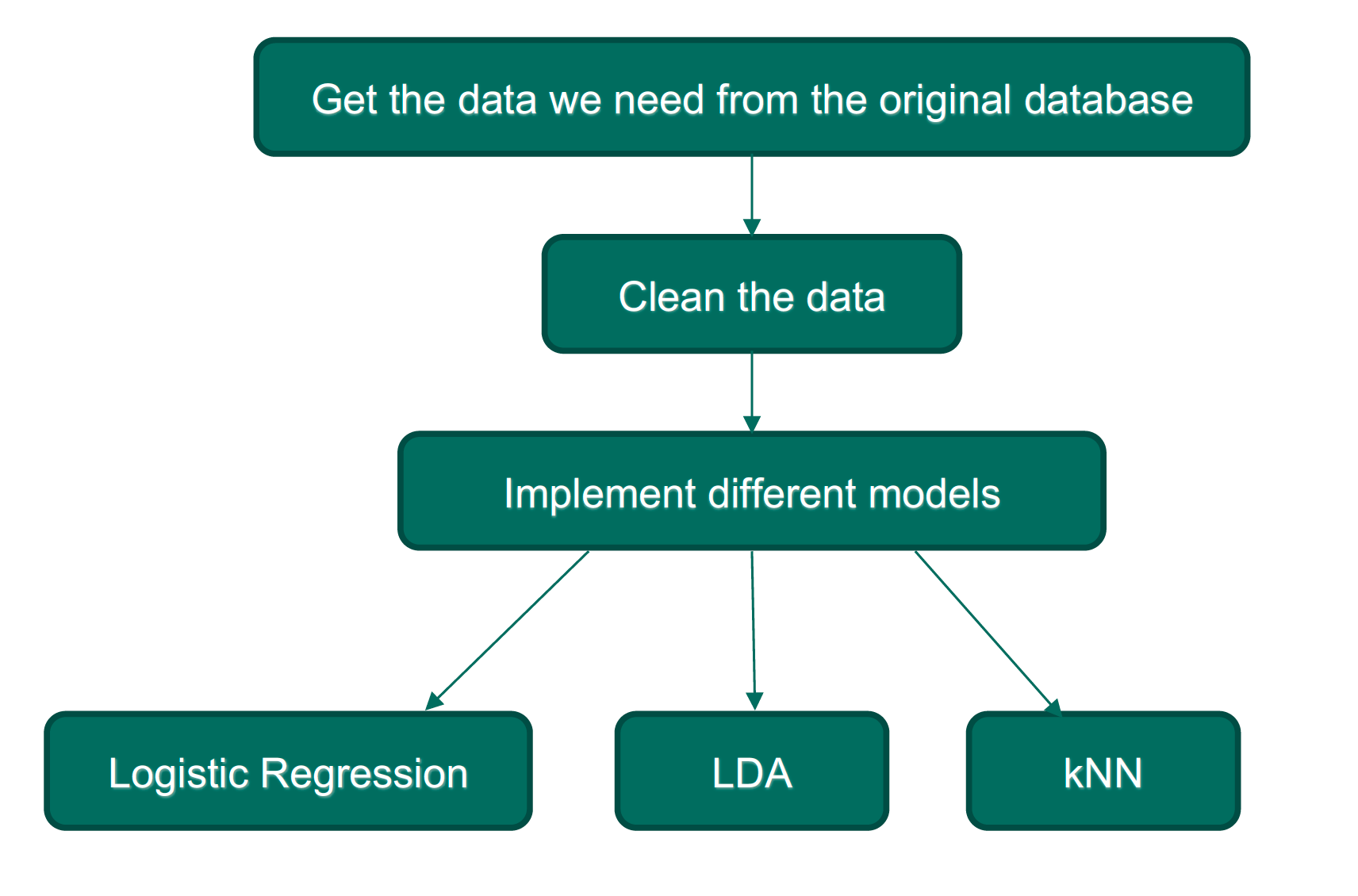
Health is a man's(or women's) most important asset; without a healthy body it would bring a lot of inconvenience to his/her daily life, especially if it's a long term health issue that plagues you for years. Dietary management is also another widely discussed topic in recent years, with higher average health and life standards. People have started to make more emphasis on what we consume on a daily basis, let that be for a better posture, to enjoy food better, or to just get a healthier body to support that entity on the long run. Diabetes in particular is a strong and more recognizable disease, with its unique characteristics such as being extremely hard to identify, extremely common especially among the more elderly, and being strongly connected with one's diet lifestyle.

## Problem Statement :

As diabetes being an hard to recognize disease, we intend to use machine learning methods to analyze and predict possible victims of the disease, and maybe even give us an indication of what are some main causes of diabetes. As we understand that diabetes is strongly connected to one's dietary habits, we intend to use one's dietary habits as the main features to discover whether diets really have such correlation with diabetes, and if yes, what are some diets we should avoid or consume more often and frequently.

Language: Python

## Methods :



*Brief Summary of the method we used*

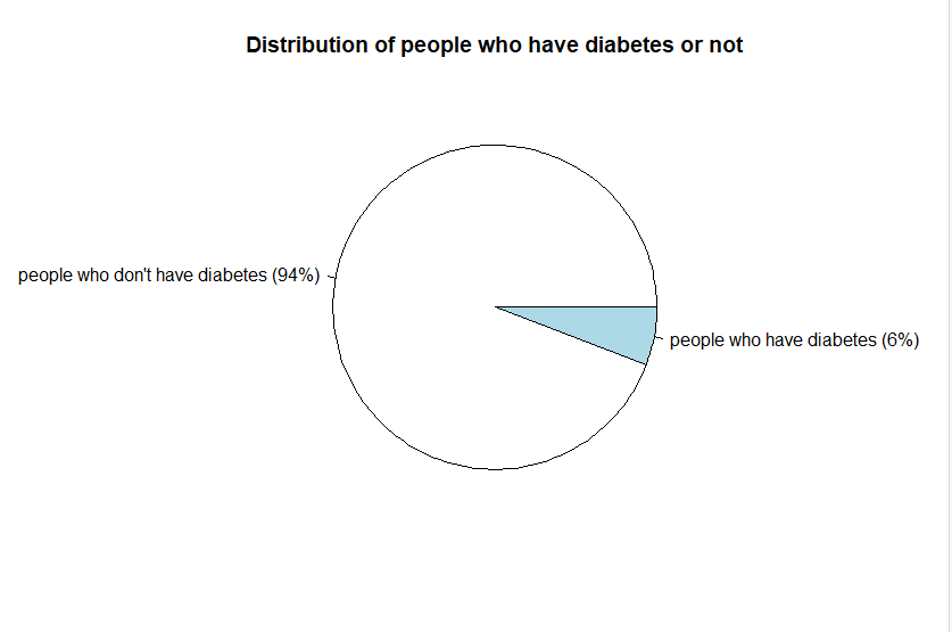
Since the dataset contains more information other than diets and disease for our prediction, the first thing we are going to do is to clear the data. After that, we implemented three different models to train our data and compared their results.

The first method we chose is *LDA(Linear Discriminant Analysis)*, which is a generalized version of *Fisher’s linear discriminant*, in other word, a linear regression model. In this case, the target is do a two-class classification.

The second method is Logistic Regression, which is a statistical model based on logistic function to model a binary dependent problem. Since in our case, the output will only have diabete/not have diabete, which is compatible with logistic regression model.

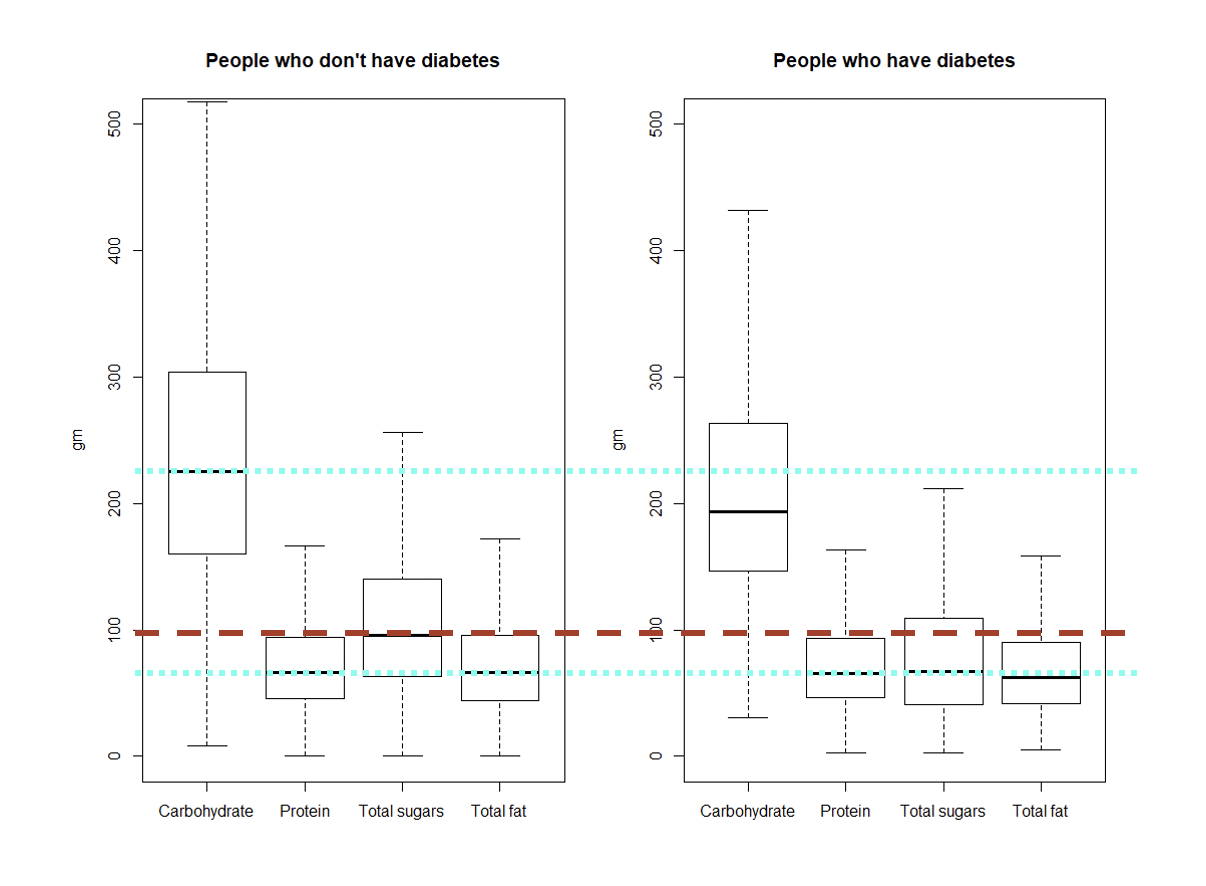
The third method is kNN(k-nearest neighbor), which is a non-parametric classification model which is based on the instances by computing the local distance for data-points.

## Preliminary Results:



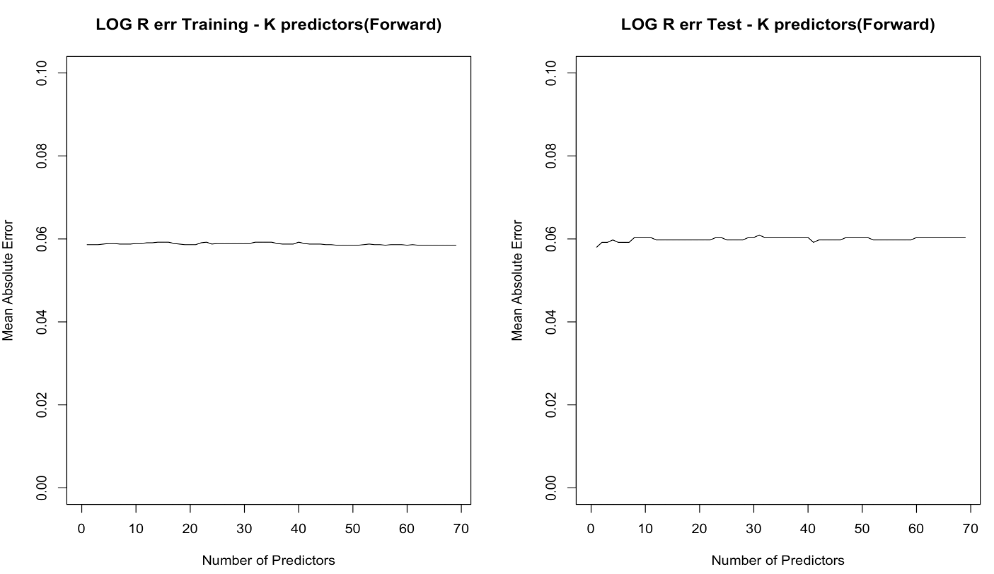
*Distribution of the cases*

The *Distribution of the cases* shows that each class label is not balanced in this dataset. 94% of positive samples. Only 6% of positive samples. To solve this problem, sampling techniques may be used in the next milestone.

**

*A brief Comparison of the differences of diets between people have/don’t have diabetes*

Before implementing the models, we firstly do a brief comparison of the difference of diet between the two groups. Some interesting data is chosen to be shown as above. As it can be seen in the picture, one obvious difference is that the people who have diabetes take clearly less sugar than the people who do not have, the same thing happens for sugar they intake. One possible reason for this may be because the people with diabetes are taking more care of the amount of sugar they take-in since their bodies are sensitive to carbohydrates.

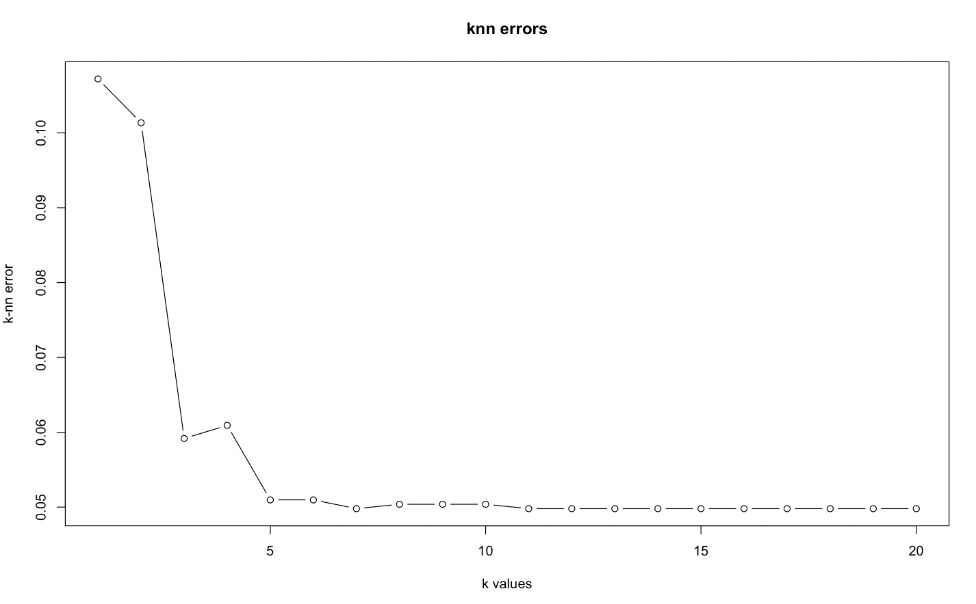


*Error rate VS Number of Predictors when applying Logistic Regression*

**

*Error rate VS Number of Predictors when applying LDA(Linear Discriminant Analysis)*

With the growth of the number of predictors, the error rate does not change much in both Logistic Regression and Linear Discriminant Analysis models. One possibility is that the gap of pf sample size between the two outputs.

**

*Error rate VS k values*

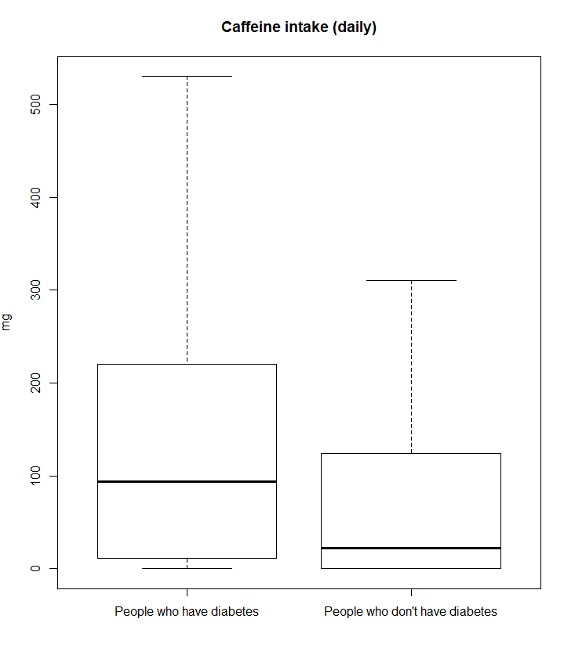
Starting from k value = 5, the error rate stabilized. When k value = 7, the model has the lowest error rate.

|  |  |  |
| --- | --- | --- |
| Method | Setting | Test Error |
| Logistic Regression | P = 1 | 5.80% |
| LDA | P = 1,2,3 | 5.97% |
| kNN | K = 7 | 4.98% |

*Error comparison of three different methods*

Comparing the three methods we implemented, the best result came from kNN with only 4.98% of test error, but the difference is not as big as we assume it to be. However, these low error rating may not be a good sign for prediction, since the scale bias between two groups are so large(16:1).

Brief conclusion:



*Caffeine intake and its relevance to diabetes*

Based on the models of prediction, we notice a really low error rate for all of three different models. However the amount of coffee intake between diabetes victims and non-diabetes are pretty significant. The mean intake of diabetes victims lands at around 100mg, however for non-diabetes people, it lands at an astonishing 25mg per day. Caffeine intake is one of the most affecting features in the prediction of the possibility of diabetes for a person.

## Dataset:

**National Health and Nutrition Examination Survey**

https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey

## Plan:

After careful scrutiny, we noticed that the dataset has very few positive cases, to elaborate: the data mostly consists of negative cases and strongly affects our model's accuracy and therefore would require better models to solve this issue. We intend to implement other common models in the future and even different sets of feature selection to improve our accuracy and narrow down the cause of diabetes. Another possible implementation would be to discover the relationship between dietary habits and other diseases, take for example heart disease, cancer, or other cardiovascular disease.