# **Machine Learning Group 70 Project Report**

## **Team Members**

YIFAN PEI 20303063

MIN WU 20302508

## **Motivation**

Several years before, with the quality people’s lives is becoming better and better, we have started to think about how to live healthier, during these years, obesity has become one of the most serious healthy questions for all humankind, and one of the factors is daily food, so we want to find the possible general pattern beneath the situation, so that we may find a reference to follow in solving obesity question.

In our project, we will use three data sets, which describe different people’s daily food proportion. For example, in “Fat\_Supply\_Quantity\_Data.csv”, this data set describes the fat content in people’s daily food from different countries. The screenshot of this data set showed as below.

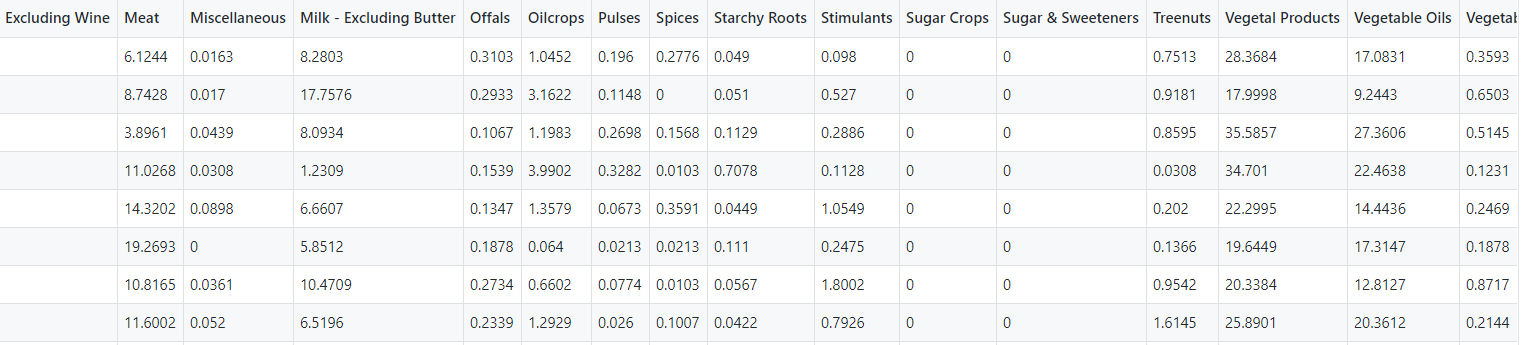


Fig1 Part of data sets (a)

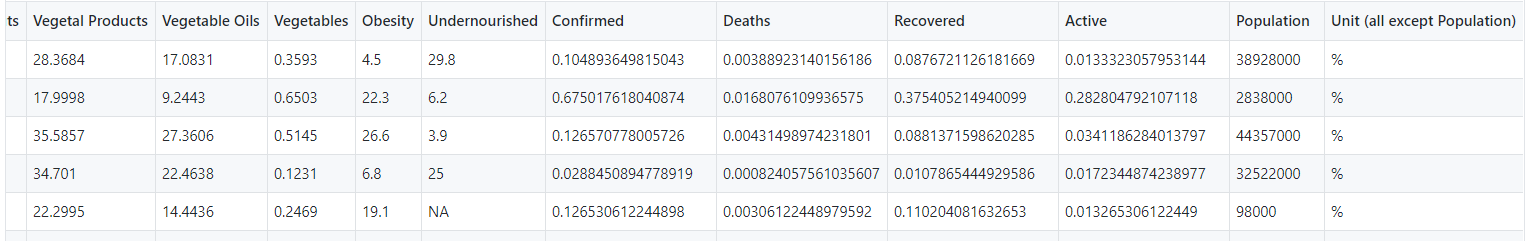


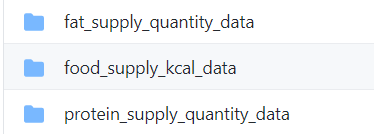
Fig 2 Part of data sets (b)

So basically, our input is a series statistics data of different daily food composition proportion in different countries’ people, and we are using one of three models (Random Forrest, Ridge and Linear Regression) to predict the probability of obesity.

## **Dataset**

We found our data sets in public data collections website.

There are three csv files as our whole data sets.



There is the same column in each data set, the only difference is that the dimension of describing food is different. For example, in “fat\_supply\_quantity\_data”, the data file looks like:





To be more precise, the different column in each data set is different element in food, like “alcoholic product”,” vegetable products”, all these columns represent how does the proportion they have in daily food in a country averagely.

In each data file, there are 171 row data.

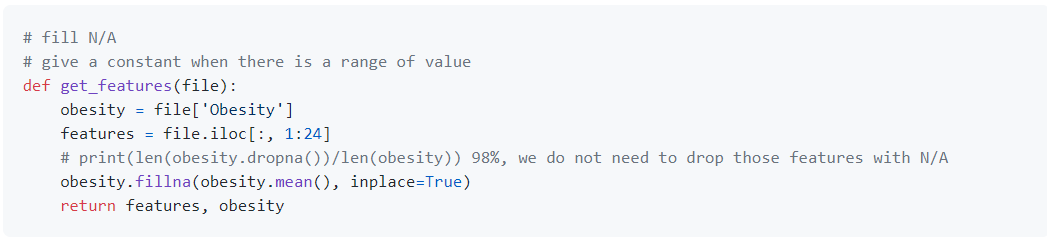
Comparing with the amount of the whole data, we think we should use “Polynomial Features” to increase the amount of training data in case over-fitting. All the data’s unit is percentage, which means the real value of them is between 0 and 1, so we decided that we do not have to do data normalization.

Since we want to find the possible relation between food element and obesity, basically we do not want to consider “country” this series data, for example some Islamic countries, they have their own religious taboo which has some influences on their food habit, which will lead the statistics data different with other religion countries, we want this kind of difference exists because we want to treat all the data equally, so we will not choose “country” this column.

Besides, we also find that in some data sets, there will be value missing, for example in “fat\_supply\_quantity\_data”, the values from “Obesity” are incomplete. We believed it is a serious question because “obesity” is our model’s output, so we started to search appropriate solution.

Firstly, we counted the amount of missing values’ percentage in total, and our conclusion is that only 2% data is missing so we decided to continue using the data sets rather than changing to another one.

Then we used *pandas* to fill the missing data. Generally, we decided to use statistic calculation based on the rest of data. After several tests of strategies including:” mean”, “all zero’,” min”, “max”, etc. We decided to use “mean” since it has the relatively best performance.



So, for the training data, we choose column data from “Alcoholic Beverages” to “Obesity”, there are 23 original features in total.

## **Method**

In our project, we are using three kinds of machine learning algorithm: Linear Regression Model, Ridge Model and Random Forrest Regression Model.

**Linear Regression Model**

To be honest, linear regression model is the model that we learned first in the lecture, so we want to introduce it firstly.

Basically, I think the final aim of any prediction question is to build a regression equation between input and output, and regardless of data augmentation, all the data from input totally should base on the real-world statistics data. As for regression equation, it is a concept which origins from statistics theory.

The first step is to get an initial regression equation, which can be set manually, then it is going to training, before introduce training process, I think it would be better to introduce loss function firstly.

As far as I am concerned, loss function is a method to evaluate the quality of the regression equation, and if it is in the leaning process, loss function can access the quality of learning. Using Gradient Descent to find the point(s) which could lead the value of loss function smallest, and it is the end of learning process (if programmer does not set ending rules additionally), and at this point, we can say that the model has a relatively best performance on training data.

**Ridge Model**

We guess it is very appropriate for us to introduce ridge model after the introduction of linear regression model.

**Random Forrest Regression Model**

To discuss random forest regression model, we must start up with random forest algorithm.

## **Experiments/Results/Discussion**

## **Summary**

## **Contribution**

At the very beginning, we were discussing the proposal of our project, and we did not split that job into several parts, but we used WeChat to have a real-time meeting to discuss our proposal.

When we are going to the next step, we divide the work into two parts: the actual coding and documentation, I am responsible for the part of coding and WU MIN is responsible for the documentation.

Then we are discussing the basic structure of our document, wrote the titles, etc. Then we are going to do what we should do.

I finish the core coding, including data pre-processing, model establishment and training, hyperparameters’ experiments and selection.

For WU MIN, she wrote the whole document, and she also examine my code, results, and our conclusion.

## **Project Link**

All the documents, pictures, data sets and code can be found in the GitHub repository, here is the link below.

PS: the report document is just a structure for our team, there is no actual content.

https://github.com/pyfppp/ML\_Group70\_Project.git