# **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.

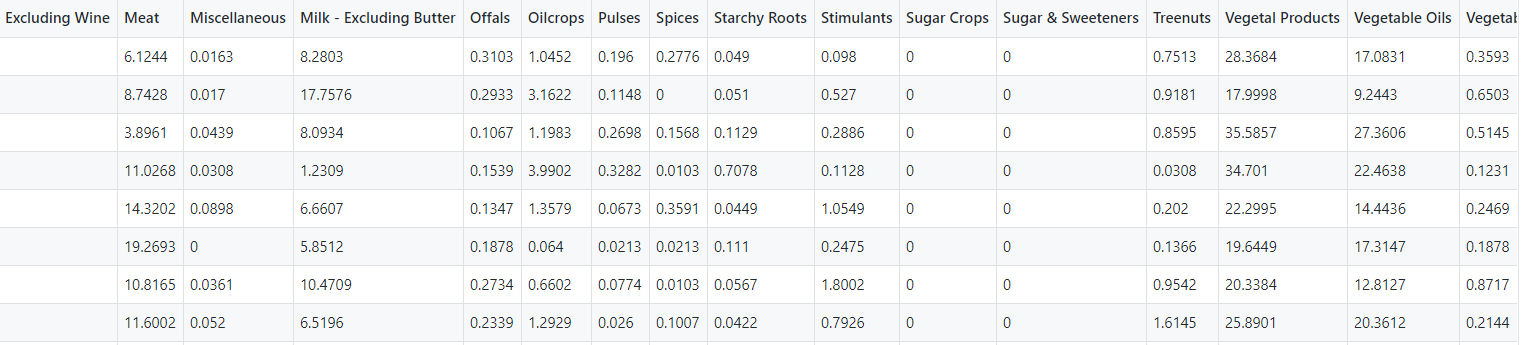


Fig1 Part of data sets (a)

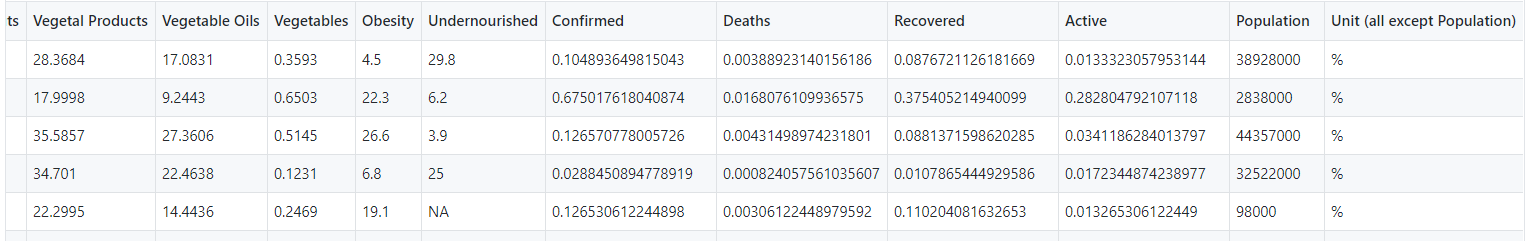


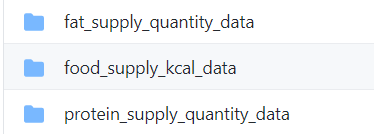
Fig 2 Part of data sets (b)

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.

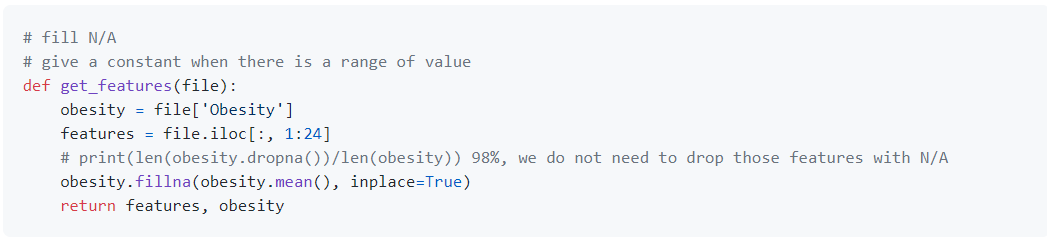
Comparing with the amount of the whole data, we think we should use “Polynomial Features” to increase the amount of training data in case over-fitting. All the data’s unit is percentage, which means the real value of them is between 0 and 1, so we decided that we do not have to do data normalization.

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.

In every machine learning model, we will face the same questions: over-fitting and under-fitting, there are also several means to solve or moderate this negative effect, and one of them is adding regularization item.

There are plenty ways to create regularization item, and there two kinds of commonly used regularization item, which are: L1 regularization item and L2 regularization item. And L2 is used in Ridge model. Basically, ridge model is a kind of upgraded version of classic linear regression model which use L2 regularization item to try to solve over-fitting problem.

**Random Forrest Regression Model**

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

To explain Random Forrest Regression model, introduction of decision trees.

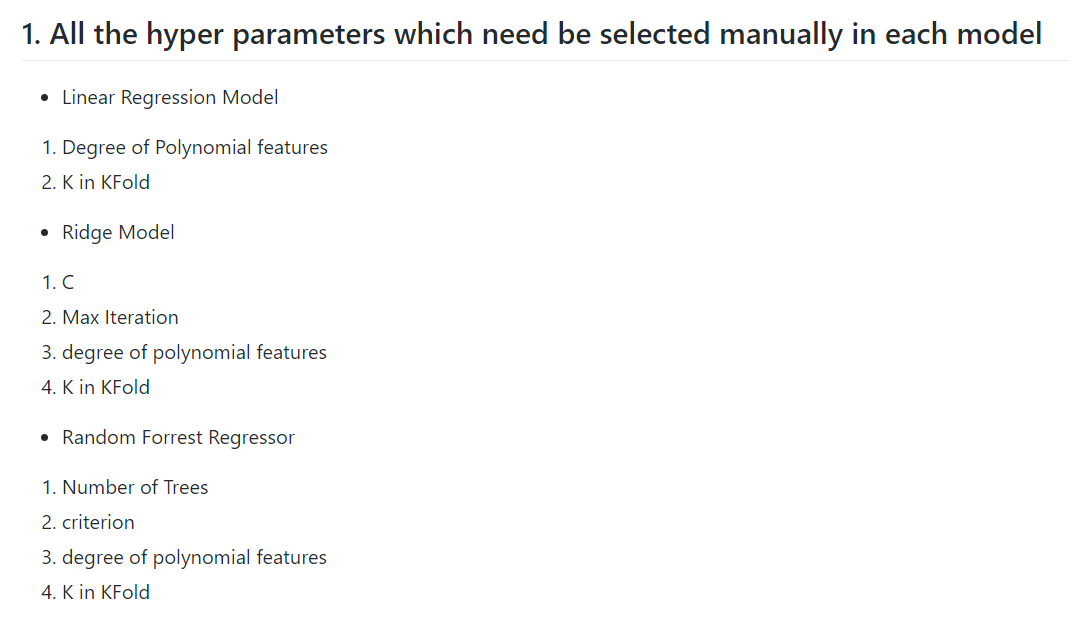
Decision trees algorithm in machine learning is a kind of decision support tool which uses tress’ structure. It is also a very effective method in many other areas like operational research. Basically, tress decision tries to split feature space into independent and undividable unit and calculate conditional probability for the unit in tree.

Random Forrest algorithm based on decision tree and basically, it is an application of decision trees multiply and make decision (including classification and prediction) based on the statistics results of each decision tree.

## **Experiments/Results/Discussion**

In the experiments stage, I basically did two jobs: selecting hyperparameters and training models based on the best hyperparameters I got.

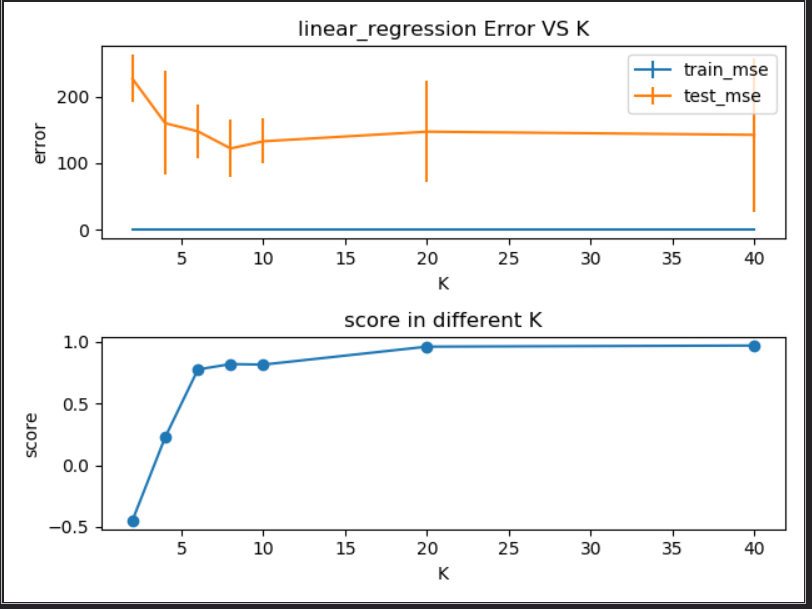
I listed all the hyperparameters I need select with cross validation method.



Basically, I need to select 10 hyperparameters in all three kinds of models in each dataset.

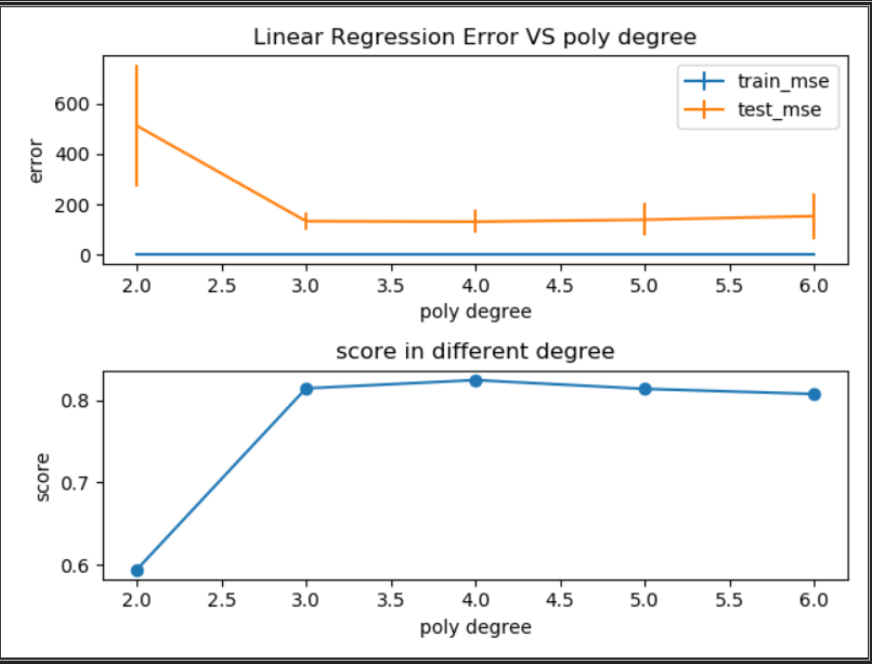
I will take an example in the first data set:” fat\_supply\_quantity\_data”.

The first is linear regression model, I use default value in degree of polynomial features to select the best K in KFold.



According to the result of cross validation experiment, I found that when degree = 8 the test error is the smallest, but the score is not the biggest, so made a trade-off here, selected 8 as the best K in KFold.

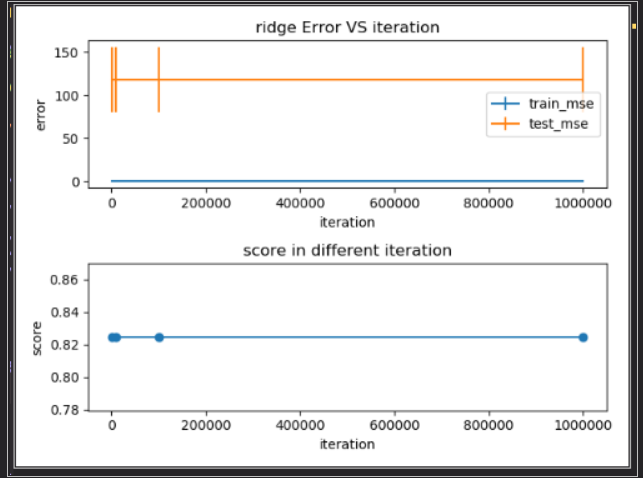
Then based on the best K value, I started to search the best degree of polynomial features, the result graph is as below.



According to the graph below, I chose the best degree as 3.

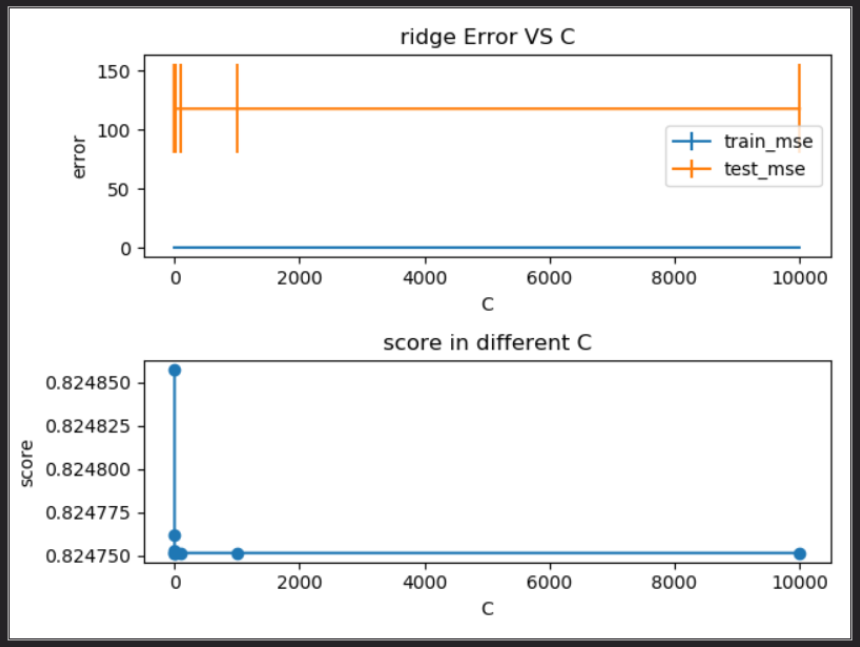
Then for the second kind of model, Ridge model. Firstly, using the same method to get the best K in KFold and degree of polynomial features, secondly, I started to search the best C and max\_iteration. The reasons why I want to adjust the max\_iteration but not using the default value is that I considered there may exist problems when the amount of whole data is insufficient and lead under-fitting.

To select C, I get the result graph as below.



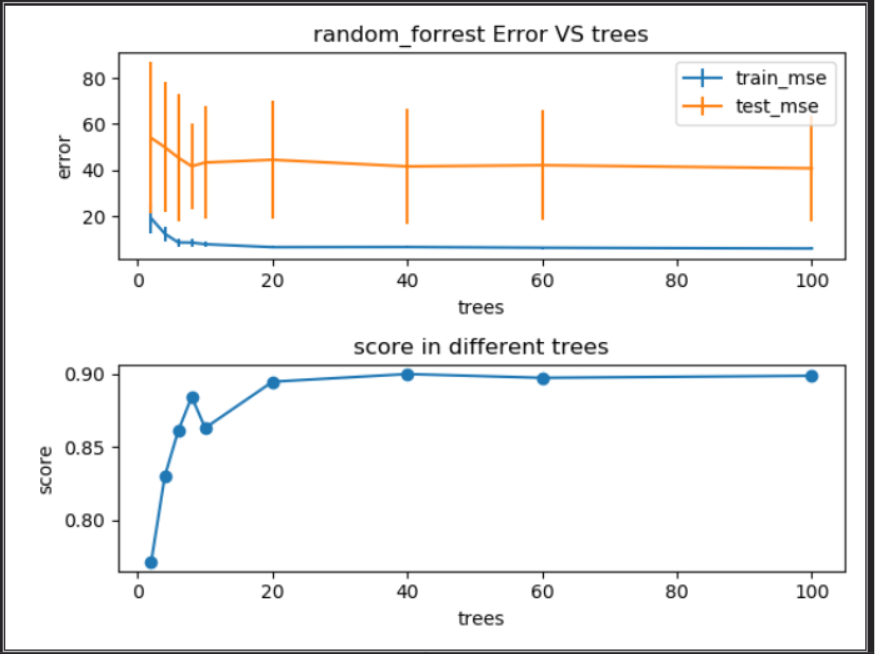
I found that in this graph I can not see there was obvious differences in test and train error due to the changes of iteration. So as a conclusion, I will use default value of max\_iteration.

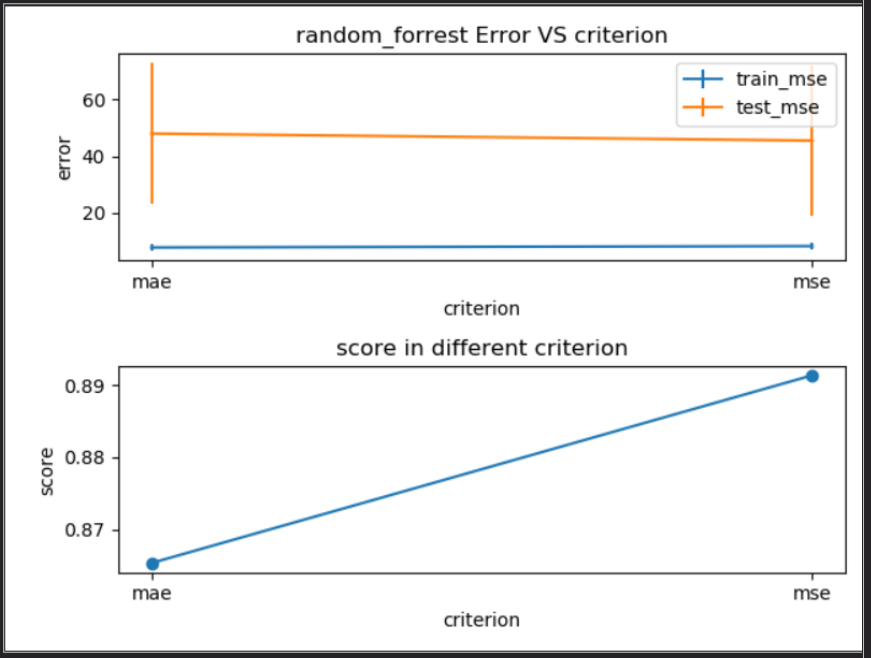
Then is C.



I found that with the changes of C, there is not obvious differences in model error (optimization range is in 1e4), I also adjusted the range of C values, but the result remained. So, I choose the default value of C as the best C value.

Then is Random Forrest Regression model, I need to select “number of trees” and “criterion” besides K and degree.





Here I want to mention that “criterion” means the rule of measuring loss.

Basically, the primary metrics are: MSE error and model score.

## **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