Learning Predictive Models of Meal Behaviour

Nishtha Kalra

Submitted for the Degree of Master of Science in

Machine Learning

Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK

August 26, 2019

Declaration

This report has been prepared on the basis of my own work. Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count:

Student Name:

Date of Submission:

Signature:

Abstract

Your abstract goes here.

Contents

1 Introduction **1**
- 1.1 Aim . . . 1
- 1.2 Motivation . . . 1
- 1.3 How will this help in my future career . . . 1

2 Background Research **2**
- 2.1 Diabetes . . . 2
  - 2.1.1 Types of Diabetes . . . 3
- 2.2 Treatment of Type 1 Diabetes . . . 4
  - 2.2.1 Open Loop . . . 4
  - 2.2.2 Closed Loop - Artificial Pancreas . . . 5
- 2.3 Carbohydrates and Diabetes . . . 6
  - 2.3.1 Fiber-rich carbs . . . 6
  - 2.3.2 Starchy carbs . . . 6
  - 2.3.3 Refined sugary carbs . . . 6

3 Machine Learning Methods Used **8**
- 3.1 Supervised Learning . . . 8
- 3.2 Underfitting and Overfitting . . . 8
- 3.3 Linear Regression . . . 10
  - 3.3.1 Least Squares . . . 11
  - 3.3.2 Linear Regression in scikit-learn . . . 11
- 3.4 Ridge Regression . . . 11
  - 3.4.1 Ridge Regression in scikit-learn . . . 12
- 3.5 Lasso Regression . . . 12
  - 3.5.1 Lasso Regression in scikit-learn . . . 13
- 3.6 Random Forest . . . 13
- 3.7 Unsupervised learning . . . 14
- 3.8 Clustering . . . 15
- 3.9 K-means clustering . . . 15
  - 3.9.1 Silhouette . . . 16
  - 3.9.2 K-means clustering in scikit-learn . . . 17
- 3.10 Moving Average Smoothing . . . 17
  - 3.10.1 Centered Moving Average . . . 17
  - 3.10.2 Trailing Moving Average . . . 17

4 Data **19**
4.1 Relation of Carbohydrates with time 20
4.2 Relation of Carbohydrates with day of the week 21
4.3 Relation of Carbohydrates with Language respondent used . 23
4.4 Relation of Carbohydrates with food type 24
4.5 Relation of Carbohydrates with eating occasion 26
4.6 Relation of Carbohydrates with Did you eat this meal at home 27
4.7 Correlation matrix and graph 28

5 Solution Methods **29**
5.1 With many features 29
5.1.1 Procedure 29
5.1.2 Results 30
5.2 Clustering and Linear Regression 31
5.3 With Moving Average 31
5.4 Future work 32
5.4.1 Robust one 32
5.4.2 LSTM 32

6 Professional Issues **33**

7 Self Assessment **34**

8 AI Ethical **35**

9 How to Use my Project **36**

References **37**

1 Introduction

1.1 Aim

1.2 Motivation

1.3 How will this help in my future career

2 Background Research

2.1 Diabetes

Living organisms require energy to power up their systems. Energy is needed by our bodies to do any simple task ranging from digestion and absorption, exercise, work, play, eat. Even activity as simple as sleeping needs energy. Where does this energy come from? The food we eat is responsible for providing us with this energy. This food provides us with a form of sugar known as glucose. Glucose is the most crucial element needed in the primary health system. The cells in our body are powered by glucose, which gives us the energy to achieve basic tasks of our everyday life.
Glucose mainly comes from food rich in carbohydrates like bread, potatoes etc. As we are eating food, it travels down our oesophagus into our stomach. While it is getting digested, our acid and enzymes present in our stomach break down the carbohydrates from the food to make glucose. Then this glucose goes into our intestines where it gets absorbed. From there, it passes in our blood to reach our cells. As this glucose is travelling from your bloodstream to your cells, this glucose is known as blood sugar.

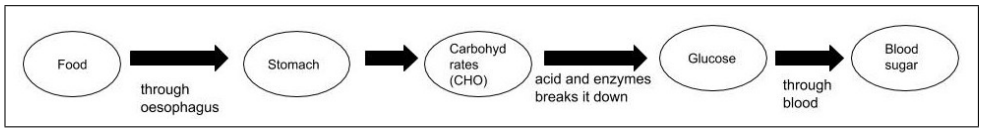

Figure 1: How Glucose is generated in our body

Now, glucose cant directly reach our cells through blood. It needs something known as insulin. Insulin is a hormone produced by the pancreas. Insulin is like a key to the door of our cells that lets glucose enter. In other words, insulin unlocks muscle, fat and liver cells so glucose can get inside them. Our body keeps the level of glucose in our body constant. The blood sugar level in our body is monitored every few seconds by beta cells present in our pancreas. When we eat carbohydrate-rich food, the blood sugar rises, thats when beta cells, present in the pancreas, releases insulin in our bloodstream.

The glucose that is not used by the body is stored in the liver in the form of glycogen. The amount of glycogen stored in the liver is enough to give energy to our body for one day. Our blood glucose level drops if we dont eat for some hours. There is no insulin produced by the pancreas. Thats when a different hormone known as glucagon is produced by the pancreas that lets the liver to break down glycogen into glucose.

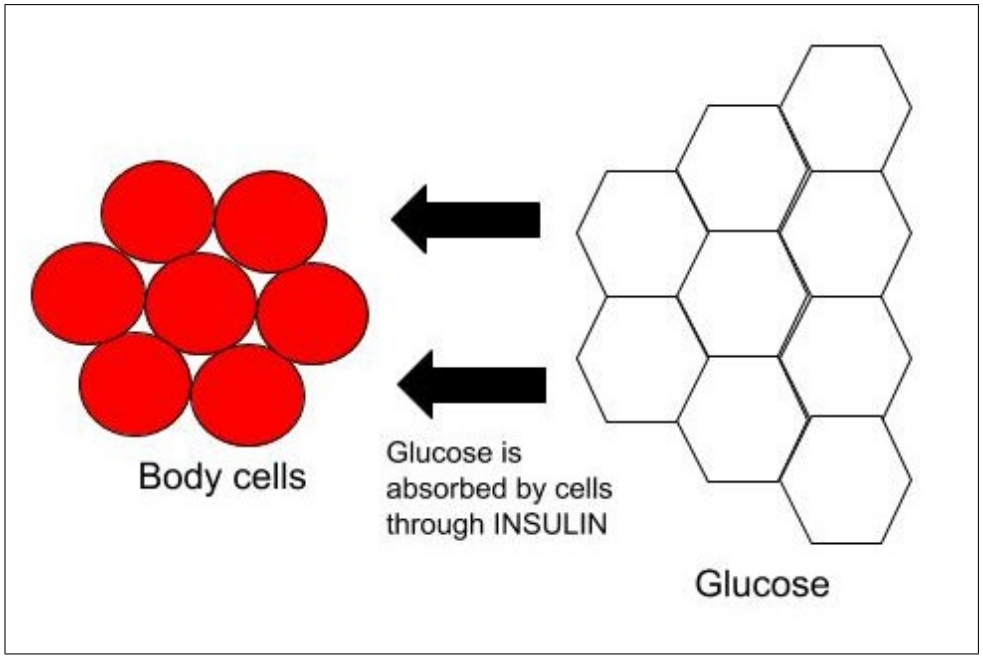

Figure 2: Purpose of Insulin in our body

Diabetes is a condition when our body is unable to break down glucose into energy [17]. The blood sugar level in a persons body is too high. This is because either the insulin produced by the pancreas is not able to work correctly or the insulin produced is not enough. This results in sugar getting accumulated in our blood. Symptoms include frequent urination, loss in weight, feeling tired, all-time thirsty, several infections, wounds dont get healed quickly etc. ultimately damaging our heart, eyes, kidney and feet. When we dont have diabetes, the pancreas makes the correct decision as to when and how insulin needs to be released that can help absorb glucose. But this does not work if we have diabetes.

2.1.1 Types of Diabetes

There are two types of diabetes, Type 1 and Type 2.
Type 1 Diabetes: 1 out of 10 people who have diabetes has Type 1 diabetes. In Type 1 diabetes, insulin is not prepared by the pancreas. Our body attacks the cells responsible for producing insulin; therefore, our body cannot produce it. When there is not any insulin to absorb the glucose present in our blood, more and more glucose gets build up in our blood. Our body

tries to get rid of the glucose via kidneys through urination, which results in frequent urine and an extreme feeling of thirst. Our body ultimately feels exhausted because it cant get the energy from glucose. Our body compensates for this loss of glucose by breaking down fats stored in our body and using them for energy. This results in loss of weight.
In healthy people, blood sugar values range between 70-200 mg/dL [14]. But in Type 1, the blood sugar level is very high. According to Diabetes UK, 4.7 million people in the UK have diabetes. 8% have Type 1 Diabetes, 90% are having Type 2, and the remaining 2% have a rarer type of diabetes. We are going to delve down more about the treatment of Type 1 diabetes later.
Type 2 Diabetes: In Type 2 Diabetes, the insulin produced by your pancreas either cannot work properly, or the pancreas cannot make enough insulin. Common synonyms include feeling tried, extreme thirst, frequent urination.

2.2 Treatment of Type 1 Diabetes

Type 1 Diabetes can be dealt with when the patient is provided with insulin, but care needs to be taken to avoid hyper and hypoglycemic episodes. Hyperglycemia occurs when blood sugar levels are too high (140mg/dL). People develop hyperglycemia if their diabetes is not treated correctly. Hypoglycemia sets in when blood sugar levels are too low (60mg/dL). This is usually a side effect of treatment with blood-sugar-lowering medication [11].

2.2.1 Open Loop

In the open-loop method, the patient injects insulin to him/herself at different times of a day. They usually inject insulin in the morning to provide the basal insulin requirement throughout the day. Basal insulin, also known as background insulin, keeps the blood glucose level consistent during fasting. In fasting, the body evenly releases glucose into our blood. Basal insulin is used to keep blood glucose level under control such that the cells can absorb glucose for energy. It is usually taken once or twice in a day. Once injected, it can provide a steady release of insulin all day.
On the other hand, another type of insulin, known as bolus insulin, is specifically taken at mealtimes to keep blood glucose level under control after a meal. Bolus insulin acts quickly on our body. Bolus insulin is usually taken before meals. In some cases, people take bolus insulin during or just after a meal to prevent hypoglycemia. The amount of insulin to inject will depend on both a measurement of glucose and on an estimate of the amount of food that is about to be eaten.

2.2.2 Closed Loop - Artificial Pancreas

Closed-loop insulin delivery is an emerging technology helpful for people who have Type 1 diabetes. It is a device consisting of a continuous glucose monitor, a control algorithm and an insulin pump [14]. The continuous glucose monitor provides glucose measurements after a regular period to the control algorithm. This algorithm is responsible for maintaining healthy blood glucose levels to avoid hyper and hypoglycemia. This control algorithm is running inside the insulin pump. These components together act as a device which regulates insulin intake inside a patients body, thus the name artificial pancreas. Wireless communication facilities automate data transfer between components. But the critical element of the artificial pancreas is the control algorithm [8].
A wearable artificial pancreas closes the loop between a glucose sensor and an insulin infusion pump. This significantly improves the quality of life of diabetic individuals. The involvement of the patient in maintaining glucose control is minimal [9]. Such a system would be able to determine the insulin requirement in real-time, regardless of the situation, and deliver the proper insulin dosage. It would be able to change the infusion as the patients activity changes and, ideally, would exist internally, eliminating the requirement of wearing external equipment. Such a system would also aim to significantly reduce the number of injections required or to eliminate them.

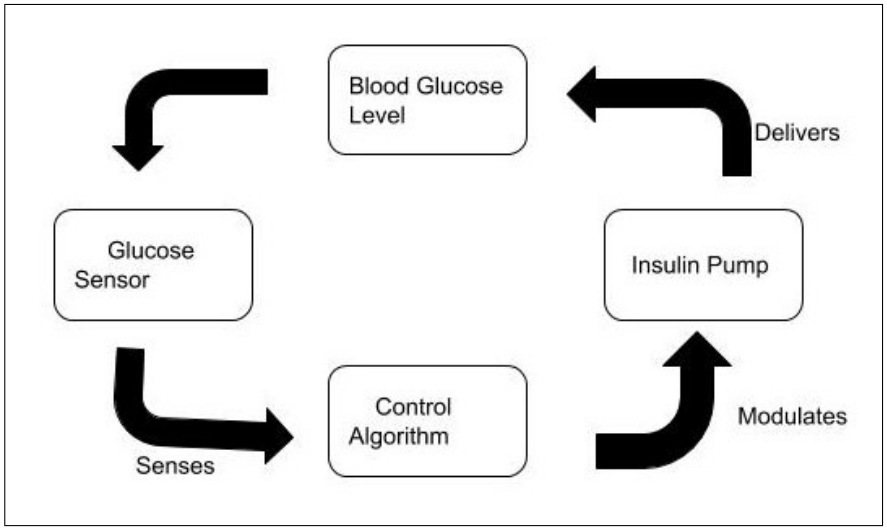

Figure 3: Artificial Pancreas

2.3 Carbohydrates and Diabetes

Types of carbs To put this in perspective, consider the three main types of carb foods:

2.3.1 Fiber-rich carbs

This includes vegetables, fruits and legumes. These foods are absorbed slowly because of their high fiber content and will thus help control blood sugar and hunger. These foods are loaded with nutrients, promote health and are calorie-dilute. Having these foods at the base of your food pyramid is a great idea. Eat them anytime.

2.3.2 Starchy carbs

Quinoa, amaranth, sprouted grain breads, potatoes, yams, acorn squash, oats, sprouted grain pasta, cereals, and similar foods are very dense sources of carbs. They are a bit lower in nutrients than the fiber-rich carb foods.
These types of starchy carbs are best consumed after exercise. During this time, your muscles are like a big sponge and will use the carbs efficiently. Consume these starchy carbs during the 3 hours or so after exercise.
Remember, energy balance is still important: Keep portion size moderate. Generally a serving is about the size of your fist. Thats a good place to start. Note: Outside of the 3 hour post-workout window, having a 1/4 cup of sweet potato or wild rice for dinner isnt going to be a huge carb load for the body to deal with. If you can meet your compliance goals and keep good eating habits with small amounts of starchy carbs, then go ahead. But be aware: the slope can get slippery. 1/4 cup can turn into a big bowl with added butter, which might mean overeating and no fat loss 3 weeks later. Use a strategy that works for you.

2.3.3 Refined sugary carbs

If you want to know what foods fall under this category, just follow around most American youth. The majority of these foods are empty calories and dont do much for health. Still, eating them during and immediately after exercise may give your body a quick energy boost and accelerate recovery.
Even then, consider the big picture: what is the food going to do for your health? What other substances are in it?
You could take advantage of refined sugary carb foods by using nutrient dense sources like dates, raisins, figs and nutrition bars. Dont assume that

because you exercised, you can eat as many refined sugary foods as you want.
Eat sugary carbs rarely, and only after exercise.

3 Machine Learning Methods Used

3.1 Supervised Learning

Supervised learning is a type of machine learning where our task is to create a learning method which will maps an output to an input, given sample inputs and outputs. The sample inputs and outputs are known as training data, which is composed of samples or inputs (X), which have their labels or outputs (y). It is a target function (f) that maps input variables (X) to an output variable (Y).

$$y = f(X) \tag{1}$$

Lets take an example of the iris dataset. The iris dataset is a classical dataset in machine learning and statistics, collected by Ronald A. Fisher [12]. A hobby botanist would like to tell the species of iris flowers that she found. She has a training set of labelled flowers. The features are the length and width of the petals, and the length and width of the sepal, all measured in centimetres. There are three possible labels (species): Setosa, Versicolor, or Virginica.
In Iris dataset example,
X = length of petal, width of petal, length of sepal, width of sepal
y = species name(either of Setosa, Versicolor or Virginica)
learning method = a function that maps X to y
The outputs in supervised machine learning problems are of two types. They can either be a discrete class label or a continuous quantity. In classification, the task is to get a mapping function from input variables (X) to discrete output variables (y). The output variables are called labels or categories, and the mapping function predicts the class for a given observation.
In the case of regression, the task is to get a mapping function from input variables (X) to a continuous output variable (y). The output variable is a real-value, such as an integer or floating-point value. A way to solve machine learning problems when there is a regression is linear regression.

3.2 Underfitting and Overfitting

Generally, when a machine learning algorithm gives poor performance, then that is either because of overfitting or underfitting.
Overfitting refers to a model that models the training data too well [5].
Bias in a learning method tells the amount of assumptions made during training the method. Variance is the estimate of the change in the target

function if different training data is used.
Overfitting happens when:

- A machine learning algorithm captures the noise of the data while training.
- The model fits the training data too well. This can be seen when the training score is too high, but the test score is low.
- The model has low bias but high variance.

Overfitting can be solved by fitting multiple models on the dataset. It can also be removed by using validation or cross-validation.
An underfitted model can neither model the training data nor generalize to new data. Underfitting happens when:

- The machine learning method used cannot capture the underlying trend of the data.
- The model used does not fit the data well.
- The model has low variance but high bias.

Underfitting can be solved by trying different machine learning methods. It can also be solved by increasing the size of the training data.

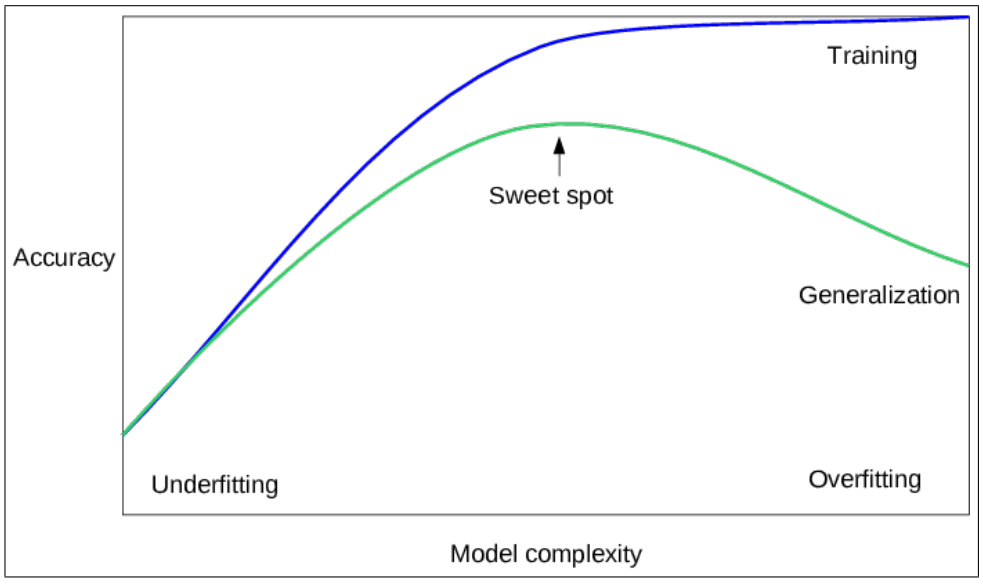

Figure 4: Balancing [19]

Figure 13 explains a balance of overfitting and underfitting in a model [19]. The blue line shows training accuracy, which increases with the complexity of the model, thus leading to overfitting. If the model is not complex, then accuracy is less, which results in underfitting. The green line shows generalization which when decreases in case of underfitting and overfitting.
Our aim for a model should be to find a good fit which is indicated as a sweet spot in the figure 13.

3.3 Linear Regression

Linear Regression is an approach to model a relationship between a dependent and an independent variable. When there are more than one independent variables, the process is known as multiple linear regression. It is represented as a linear equation that combines a specific set of input values (x) and provides a predicted solution($\hat{y}$) [3]. The input values and the output value are numeric. The linear equation is made up of coefficients and intercepts. The coefficient in a linear equation is a one case factor assigned to each input value. Intercept is an additional coefficient added to the line to give an extra degree of freedom. The intercept is also known as the bias coefficient.
The equation of the model of a simple linear regression can be written as:

$$\hat{y} = w \cdot x + b \tag{2}$$

where b = intercept and w = coefficient
Dependent variable in this equation is $\hat{y}$ and independent variable is x.
When training a regression model, the coefficients are learned and fitted to training data. The aim is to find the best fit line and minimize the cost function. We try to minimize the error between actual and predicted values. We can measure the error using the cost function [7]. Coefficients are also known as slope. In equation 1, w is the effect on ŷ when x is increased by 1 unit.
When we have more than one input, the line is a plane or a hyperplane. Suppose if we have p features or inputs then equation 1 can be written as:

$$\hat{y} = w_0 \cdot x_0 + w_1 \cdot x_1 + \ + w_{(p-1)} \cdot x_{(p-1)} + b \tag{3}$$

where where x_j is the $(j + 1)$st feature and w_j are its coefficient.

The dependent variable here is $\hat{y}$ and independent variables are x_0 to $x_{(p-1)}$. When the parameters are estimated, they are then used for prediction. Please note that in a linear model, we aim for linearity in the parameters; there does not have to be linearity in the attributes.

3.3.1 Least Squares

The parameters need to be estimated to make a prediction in any machine learning algorithm. In the case of linear regression, the parameters w(slope) and b(intercept) are estimated using the approach of Least Squares. We choose w and b such that the Residual Sum of Squares or RSS is minimized [20].

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{4}$$

where y_i is actual label and $\hat{y_i}$ is predicted label for (i)th data of size n.
The Total Sum of Squares or TSS can be calculated using the following formula:

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{5}$$

where y_i is actual label for (i)th data of size n and $\overline{y}$ is the average label:

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{6}$$

where y_i is actual label for (i)th data of size$n$.
R^2 is the measure of variability in the label. It is given by the following formula:

$$R^2 = \frac{TSS - RSS}{TSS} \tag{7}$$

where TSS and RSS is calculated using the formula given in equation (4) and (5).

3.3.2 Linear Regression in scikit-learn

3.4 Ridge Regression

In figure 12, we can move to the left of the graph by using regularization [20]. Ridge regression is a technique of regularization when there is overfitting in a model where linear regression is used. In ridge regression the coefficients are chosen such that:

1. Their magnitude is as small as possible so that their effect on prediction is minimum.

2. However, they should be good enough to predict results with reasonable accuracy.

The parameters are coefficients (w) and intercept (b). RSS calculated in equation (4) can also be written as:

$$RSS = \sum_{i=1}^{n} (y_i - w \cdot x_i - b)^2 \tag{8}$$

In Ridge regression, w and b are chosen such that the following equation can be minimized:

$$RSS + \alpha \|w\|^2 = RSS + \alpha \sum_{j=0}^{p-1} w_j^2 \tag{9}$$

where RSS is as given in equation (8),
p is number of features,
w_j is the coefficient corresponding to the feature j and
α is the regularization parameter with condition $\alpha \geq 0$.
When $\alpha = 1$ then Ridge Regression acts as Linear Regression.
As α becomes bigger, the coefficients gets smaller.
This is also known as L_2 regularization where:

$$L_2 norm = \sqrt{\sum_{j=0}^{p-1} w_j^2} \tag{10}$$

3.4.1 Ridge Regression in scikit-learn

3.5 Lasso Regression

Ridge Regression tends to make coefficients shift towards zero. Coefficients become 0 only when $\alpha = \infty$. Otherwise, the coefficients are not 0. Therefore, all the features are used. If the number of features is too large, then using all features can be a problem and might result in overfitting again.
Here is where Lasso Regression comes into the picture. In Lasso Regression, many coefficients are set to 0. In a way, we can say that Lasso performs model selection.
In Lasso regression, w and b are chosen such that the following equation can

be minimized:

$$RSS + \alpha\|w\|_1 = RSS + \alpha \sum_{j=0}^{p-1} |w_j| \tag{11}$$

where RSS is as given in equation (8),
p is number of features,
w_j is the coefficient corresponding to the feature j and
α is the regularization parameter.
This is also known as L_1 regularization where:

$$L_1 norm = \|w\|_1 = \sum_{j=0}^{p-1} |w_j| \tag{12}$$

3.5.1 Lasso Regression in scikit-learn

3.6 Random Forest

Random Forest is a machine learning approach that uses the ensemble method for prediction. In ensemble method of prediction, multiple machine learning algorithms are taken into consideration, and their predictions are combined to make a more accurate final prediction.
Random forest is an ensemble machine learning method which comprises of multiple decision trees. The trees in the random forest run in parallel, and there is no interaction between these trees. Multiple decision trees are constructed while training a random forest, and then the output is a mean prediction of all the individual predictions of the decision trees. It is used to prevent overfitting so that the model does not rely on any single feature. Figure ¡¿ shows an implementation of a random tree which combines prediction acquired by 600 decision trees and averages them to give a final prediction.

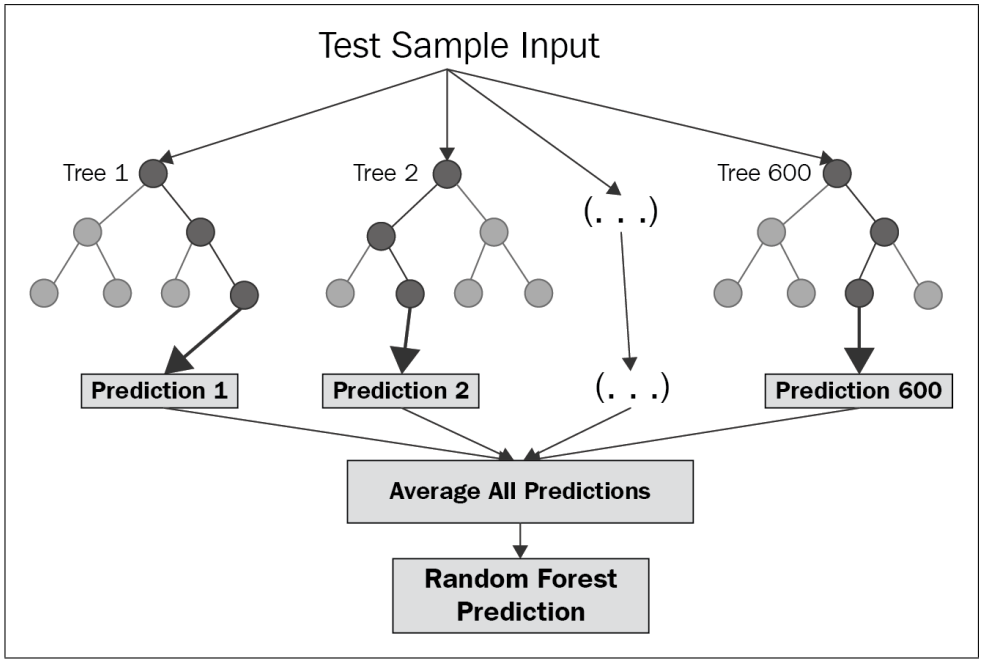

Figure 5: An example of Random Forest [13]

3.7 Unsupervised learning

Unsupervised learning is a type of machine learning algorithm which has inputs or samples, but unlike a supervised learning method, the inputs do not have the associated outputs. Therefore, unsupervised learning is used to find the underlying structure of distribution in the data to learn more about the data. They are unsupervised because there is no teacher to supervise the algorithms and tell them what is correct and what is wrong [6]. Algorithms discover the structure of data on its own without any supervision.
Let us take the example of iris dataset explained above in section (3.1). In the case of unsupervised learning, we have
X = length of petal, width of petal, length of sepal, width of sepal
But we do not have y. One way to solve this problem can be to group samples with similar features together. So the iris data with have similar lengths and widths of petals and sepals are grouped together, and we assume that each of the group or cluster is its own species.
The uses of unsupervised learning can be broadly categorised into:

- finding the underlying structure of a dataset

- grouping the data so that it can be useful
- representing data in a compressed format

The above tasks can be broken down into two methods: clustering and reducing dimensionality.

3.8 Clustering

Grouping of samples in the data in such a way that the samples that are similar to each other are grouped in a cluster is known as clustering. In unsupervised learning, clustering is used to create clusters of similar samples in a dataset. There are various ways in machine learning to create such clusters. These algorithms differ in the way they make clusters. Some algorithms consider the distance between samples as a criterion for clustering; some consider the dense area. We select the appropriate clustering algorithms and corresponding hyperparameters depending on the dataset and result desired.
Clustering is used in Biology to group genes into gene families, in social network analysis to recognise communities and many more such examples.
There are two types of clustering methods [21]:

- Parametric Clustering - density estimation with a mixture of gaussian.
- Non-Parametric Clustering - natural grouping or clustering in the dataset.

In non-parametric clustering, first, a measure of similarity or dissimilarity is defined. Then an objective function is defined which measures how well the data is represented in the clusters, and then it is optimised.

3.9 K-means clustering

K-means clustering is a type of non-parametric clustering. In K means clustering, the data is divided into K clusters or groups. When K is large, then the number of clusters is more, and then granularity in each cluster is more. When K is small, the number of clusters is less; thus, the granularity is less. In this method of clustering:
Input = the dataset
Output = a set of K labels, assigning each data of one of the label.
The method used for clustering is by calculating the centroid of each group. It can be imagined as a party where there are lots of people, and there are

some people who become the centre of attraction because they are magnetic. These people act as centroids, and other people gather around them depending on their liking.
The algorithm of K-means algorithm is as follows:

1. Randomly create K centroids for the K clusters.

2. Measure the distance of each data point with each of the K centroid. Usually, Euclidean distance is used to measure the distance. The closeness measure is a hyperparameter here.

3. Assign each data point to the cluster, which has minimum distance with its centroid. For example, if a data point has minimum distance with centroid number 2 then assign this data point a label of 2. This is given by the following equation [16]:
   If c_i is the collection of centroids in set C, then each data point x is assigned to a cluster based on

$$\min_{c_i \in C} dist(c_i, x)^2 \tag{13}$$

4. Update the centroid for each of the cluster.
   Let the centroid be S_i for ith cluster after calculating it from equation (13). Then the updated centroid of cluster i is:

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i \tag{14}$$

5. Repeat steps 2 and 4 until the algorithm converges.

3.9.1 Silhouette

Choosing the value of K or the number of cluster in K-means clustering is very crucial. An incorrect value of K might not group the data points in the dataset properly. There are some ways by which we can validate the clusters formed during K-means clustering. One such method is the Silhouette method.
The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters [15]. The value is between -1 to $+1$. If the value is high, then that means that the data point is well-matched with its cluster and poorly matched with other clusters. A negative or low value of silhouette indicates that the number of clusters is either too less

or too more. Choosing the method of distance calculation while calculating silhouette value is a hyperparameter. It can be calculated with any method like Euclidean distance of Mahattan distance.

3.9.2 K-means clustering in scikit-learn

3.10 Moving Average Smoothing

Moving average smoothing is a technique to remove variations in a time series. It is used for data preparation, feature engineering and prediction. Smoothing is used to remove noise in causal processes.
When a moving average is calculated for a time series, a new series is created which is composed of averages of raw values of original time series. A window size is initialised, which gives the number of observations that should be used to calculate the average. This window slides in the time series to calculate the average values in the new series.
There are two ways of calculating moving average namely centered moving average and trailing moving average [4].

3.10.1 Centered Moving Average

In centered moving average, the value is calculated at time t by taking into consideration the past, present and future values. For example, if the window size is three, then the value at time t is calculated as [4]:

$$ma(t) = mean(obs(t-1), obs(t), obs(t+1)) \tag{15}$$

where $ma(t)$ is the centered moving average at time t
$obs(t)$ is observation at time t
$obs(t-1)$ is observation at time $t-1$ and
$obs(t+1)$ is observation at time $t+1$.
Centered moving average method is used to remove the trend and seasonality of a time series.

3.10.2 Trailing Moving Average

In trailing moving average, the value is calculated at time t by taking into consideration the past and present values only. For example, if the window size is three, then the value at time t is calculated as [4]:

$$ma(t) = mean(obs(t-2), obs(t-1), obs(t)) \tag{16}$$

where $ma(t)$ is the trailing moving average at time t
$obs(t)$ is observation at time t
$obs(t-1)$ is observation at time $t-1$ and
$obs(t-2)$ is observation at time $t-2$. Trailing moving average method is used for time series forecasting.

4 Data

An interview was conducted by trained staff at study entry. A 24-h dietary recall questionnaire was completed by the participants, which collected the specific food items and quantities consumed by each participant from midnight to midnight on the day preceding the in-person interview [2]. All participants were asked to complete two 24-h dietary recall interviews. For both dietary recall interviews, all food items and quantities consumed by each participant from midnight to midnight on the day preceding the interview were recorded. The first dietary recall interview was collected in-person. The second interview was obtained by telephone 310 days later, although not on the same day of the week as the in-person interview. A set of measuring guides were given to participants for help in reporting food amounts during the in-person interview. These guides and a food model booklet were given to the participants to assist in reporting food amounts during the subsequent telephone interview. The 24-h recalls included both weekdays and weekend days. In this project, I have used both days for some of the solutions and only day 1 for some.

4.1 Relation of Carbohydrates with time

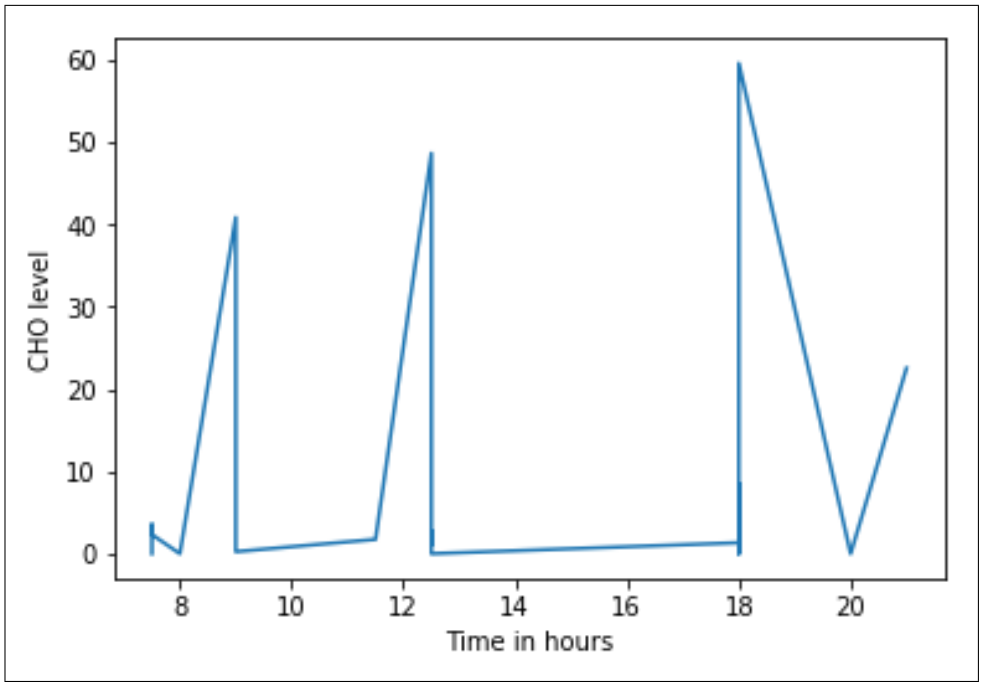

Figure 6: Time vs Carbohydrates(CHO) of one participant

I loaded the dietary interview data obtained from reference [10] in my Jupyter notebook. I created a data frame in Python Pandas with only time(DR1_020) and CHO level(DR1ICARB) of a person corresponding with the serial number(SEQN) 83732. I plotted a line graph with time(in hours) on X-axis and CHO level on y-axis. After plotting the graph, I was able to see that the carbohydrates level peaks around 9 am, 12 pm and 6 pm thus showing that carbohydrates level in our body is higher when we have breakfast, lunch and dinner. The CHO level remains almost constant between 9 am to 11 am and between 1 pm to 5 pm. Thus showing that the person did not eat anything between this period, therefore, his/her CHO level remained a constant around 0mg/dL.
The time of intake of carbohydrates plays a vital role in our diet. There have been numerous researches about when to eat which food according to its carbohydrates content. A concept known as Nutrient timing is based on this. Nutrient timing is a planned changed in food intake to promote a healthy lifestyle [1]. Nutrient timing strategies are based on how the body

handles different types of food at different times. Our body handles various types of carbohydrates differently.
Carbohydrates significantly affect the insulin response of our body. There are some carbohydrates like beans, legumes and vegetables that are high in fibre and low in simple sugar. Not much insulin is required for such food. Whereas, there is some food which has refined carbohydrates which enters the body rapidly. It can elevate blood glucose levels and bad cholesterol and can lead to insulin resistance.

4.2 Relation of Carbohydrates with day of the week

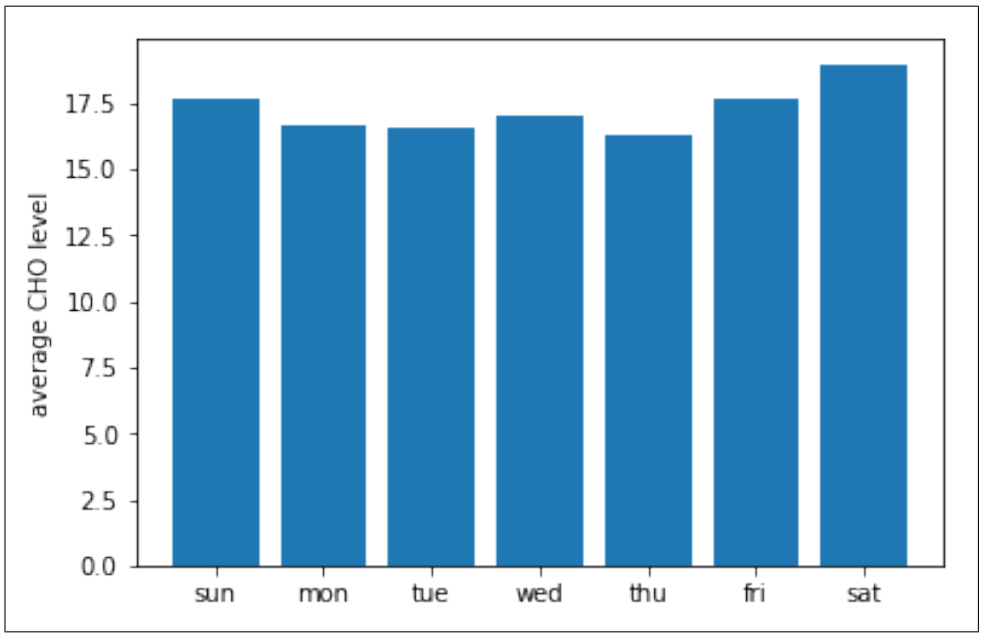

Figure 7: Bar Graph showing average CHO consumed on each day of week

I loaded the dietary interview data obtained from reference [10] in my Jupyter notebook. I created a data frame in Python Pandas with time(DR1_020), Intake day of the week(DR1DAY) and CHO level(DR1ICARB) of all people. I took an average of CHO intake of each day of the week. The codes corresponding to the day of the week are as follows:
1 - Sunday
2 - Monday

3 - Tuesday
4 - Wednesday
5 - Thursday
6 - Friday
7 - Saturday

Then, I plotted a bar graph with the day of the week and their corresponding average CHO level. After plotting the graph, I was able to see that the carbohydrates intake of people is usually more on Friday, Saturday and Sunday in comparison with the other days of the week. This suggests the partying pattern of people because people often go out for parties and dinner on weekends and usually work on weekdays. Thus, their diet contains less of junk foods on weekdays, and they prefer eating homemade and healthy food. But on weekdays, as they eat in fancy restaurants, they dont restrict themselves to healthy food only and consume food which is high in carbohydrates.

4.3 Relation of Carbohydrates with Language respondent used

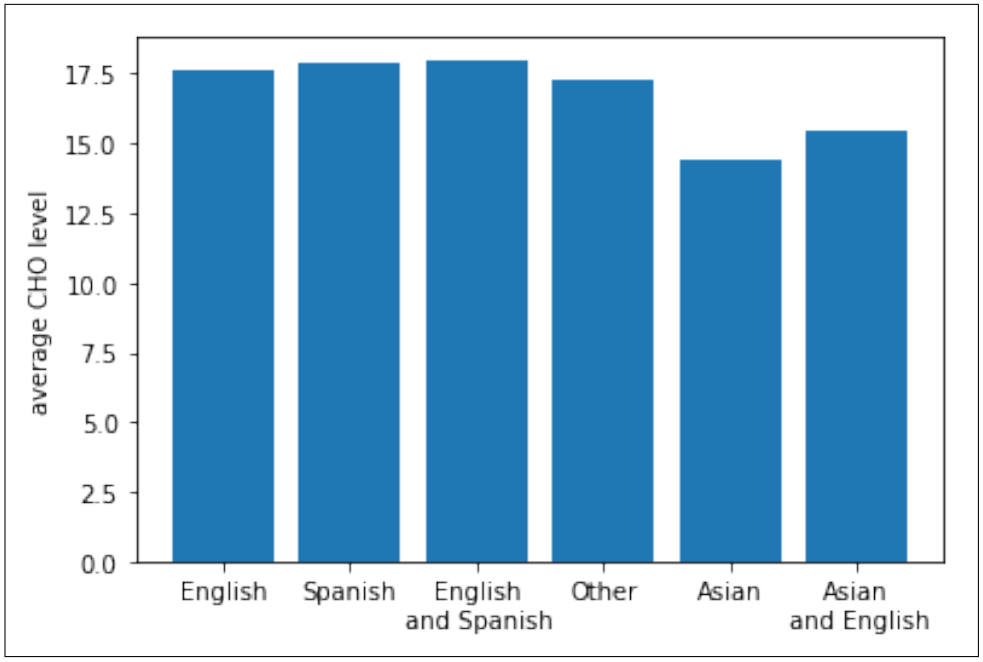

Figure 8: Bar Graph showing average CHO consumed grouped by language respondent used

4.4 Relation of Carbohydrates with food type

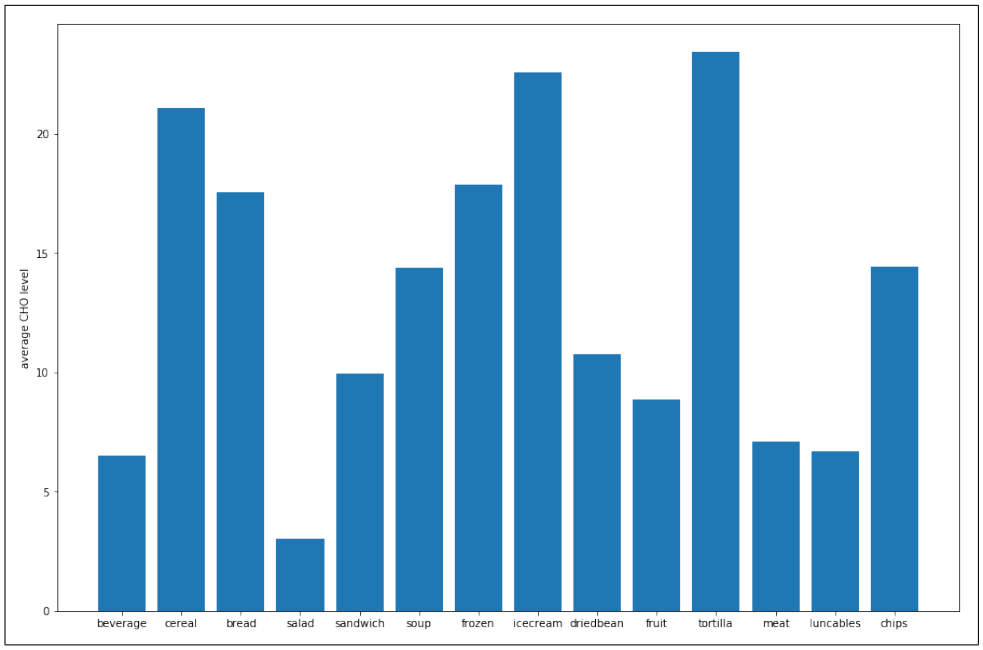

Figure 9: Bar Graph showing average CHO consumed grouped by food type

I loaded the dietary interview data obtained from reference [10] in my Jupyter notebook. I created a data frame in Python Pandas with time(DR1_020), Food type(DR1CCMTX) and CHO level(DR1ICARB) of all people. I took an average of CHO intake corresponding to each food type. The food types and their codes are as follows:
1 - Beverage
2 - Cereal
3 - Bread/baked products
4 - Salad
5 - Sandwiches
6 - Soup
7 - Frozen meals
8 - Ice cream/frozen yoghurt
9 - Dried beans and vegetable
10 - Fruit
11 - Tortilla products
12 - Meat, poultry, fish

13 - Lunchables
14 - Chips
Then, I plotted a bar graph of the food type and their corresponding average CHO level. We can see from the bar graph that tortilla, ice cream/frozen yoghurt, cereals, bread and chips are amongst the food with higher carbohydrates level. People who are trying to reduce weight try to take a less intake of such food so that the energy is obtained by the fats stored in the body, thus resulting in loss of fat for energy consumption.
According to the US Department of Agriculture [18], 100 gms of cereal has 68gm of carbohydrates, 100gm of white bread has 49gm of carbohydrates, 100gm of beer has 3.6gm of carbohydrates etc.
Please find below a table which gives a rough estimate of carbohydrates in some food types according to [18]

Food Type (per 100grams)	Carbohydrates(in grams)
Cereal	68
White bread	49
Beer	3.6
Coleslaw salad	15
Egg and cheese sandwich	18
Tomato soup	7
Frozen food	13
Ice cream	24
Beans	16.15
Fruit(Apple)	14
Tortilla	45
Meat	0
Fish	7
French Fries/Chips	41

As seen by this table, we can say that our analysis is correct. Hence, the food that we eat significantly affects carbohydrates level in our blood.

4.5 Relation of Carbohydrates with eating occasion

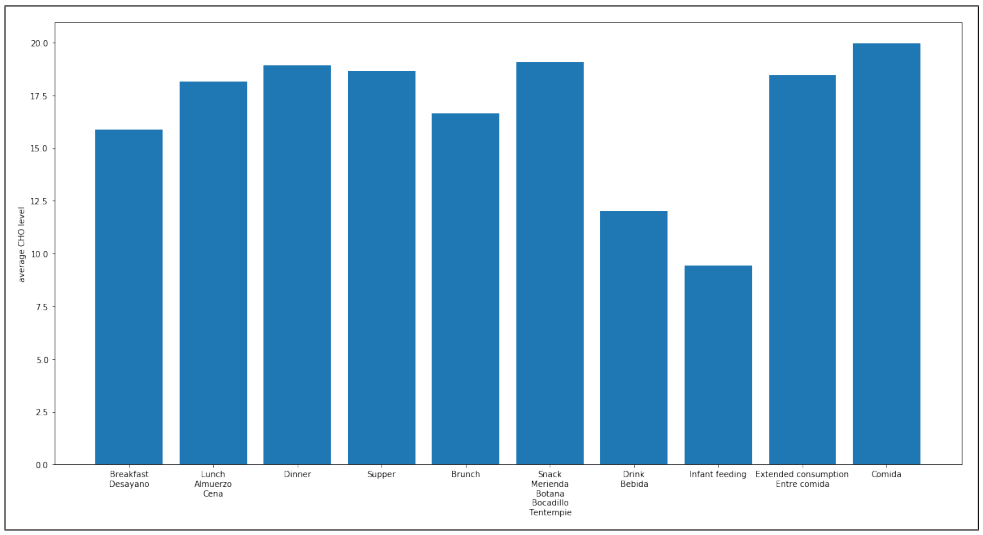

Figure 10: Bar Graph showing average CHO consumed grouped by eating occasion

I loaded the dietary interview data obtained from reference [10] in my Jupyter notebook. I created a data frame in Python Pandas with time(DR1_020), name of eating occasion(DR1_030Z) and CHO level(DR1ICARB) of all people.
As the names of eating occasion are in English and Spanish, I merged them as follows:

Eating occasion and their codes	Merged Code
1 - Breakfast and 10 - Desayano	1
2 - Lunch,11 - Almuerzo, 12 - Comida	2
3 - Dinner and 14-Cena	3
4 - Supper	4
5 - Brunch	5
6 - Snack,13 - Merienda, 16 - Botana, 17 - Bocadillo, 18- Tentempie	6
7 - Drink and 19 - Bebida	7
8 - Infant feeding	8
9 - Extended consumption and 15 - Entre Comida	9

Then I took an average of CHO intake corresponding to each occasion. I plotted the results on a bar graph with eating occasion on x axis and average

CHO corresponding to each eating occasion on y axis. As seen in figure 8, it is evident that snacking has the highest average CHO level.

4.6 Relation of Carbohydrates with Did you eat this meal at home

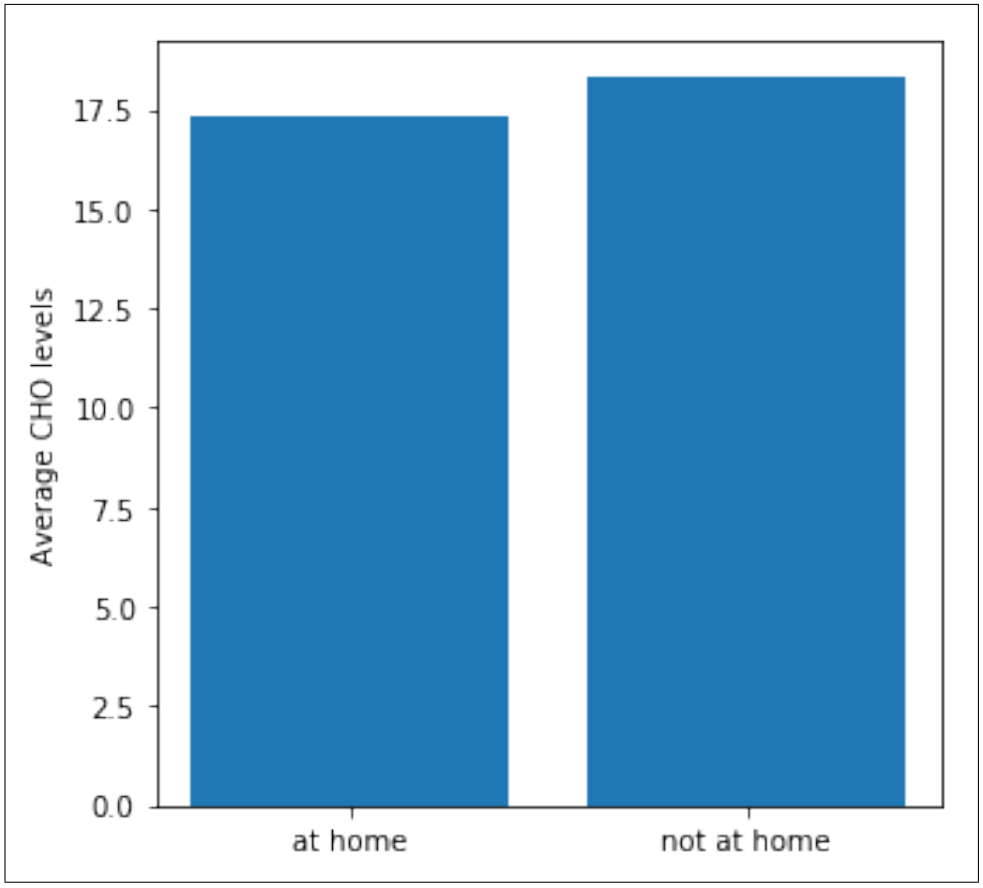

Figure 11: Bar Graph showing average CHO consumed when food is consumed at home and outside

I loaded the dietary interview data obtained from reference [10] in my Jupyter notebook. I created a data frame in Python Pandas with time(DR1_020), whether you eat this meal at home(DR1_040Z) and CHO level(DR1ICARB) of all people.
If DR1_040Z is 1 then this meal was eaten at home, if the code is 0, then

this meal was not eaten at home. I plotted a bac chart, as shown in figure 9, which indicates average CHO levels of meal consumed at home and meals not consumed at home.
The graph suggests that the meal eaten at home has less CHO levels than meal consumed outside of the home, i.e. meal not prepared at home. This can happen because usually when we cook our food at home, we are conscious about what we eat and try to reduce the amount of bad carbohydrates intake. Whereas, when we go out for meals, we are often not aware of the ingredients of our food that we order and end up eating meals that are rich in carbohydrates.

4.7 Correlation matrix and graph

5 Solution Methods

5.1 With many features

5.1.1 Procedure

- Import important libraries
- load data in pandas dataframe
- drop empty rows
- find correlation of carbohydrates with other features so as to understand which all features affects carbohydrates level the most
- The following are the features with highest effect on CHO levels:
  DR1IKCAL 0.794958
  DR1ISUGR 0.664955
  DR1IFIBE 0.569776
  DR1IVB1 0.577118
  DR1IFOLA 0.541616
  DR1IFDFE 0.519071
  DR1IMAGN 0.526465
  DR1IIRON 0.528656
- Create a new dataframe with the following values: 'DR1IKCAL', 'DR1ICARB', 'DR1ISUGR', 'DR1IFIBE', 'DR1_020', 'DR1CCMTX', 'DR1_030Z'
- convert time into hours
- Create a correlation matrix

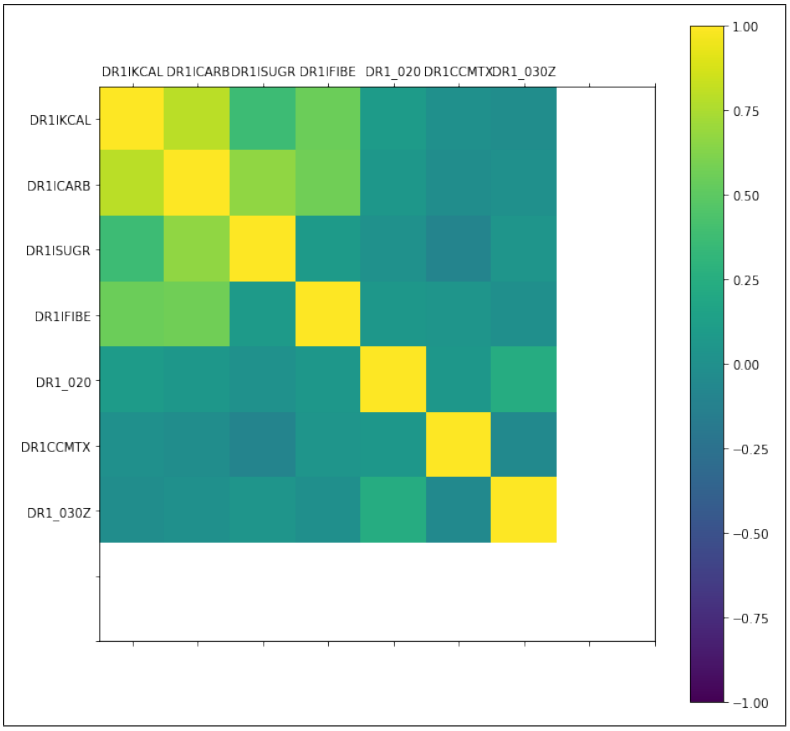

Figure 12: Coorelation Matrix

- X = 'DR1_020', 'DR1IKCAL', 'DR1ISUGR', 'DR1IFIBE', 'DR1_030Z', 'DR1CCMTX'
- y= 'DR1ICARB'
- Apply linear regression

5.1.2 Results

Score name	Score
Training set score	83.84772989545684
Test set score	84.3880020213138
R2 score	84.3880020213138
RSS score	89.22549483021835

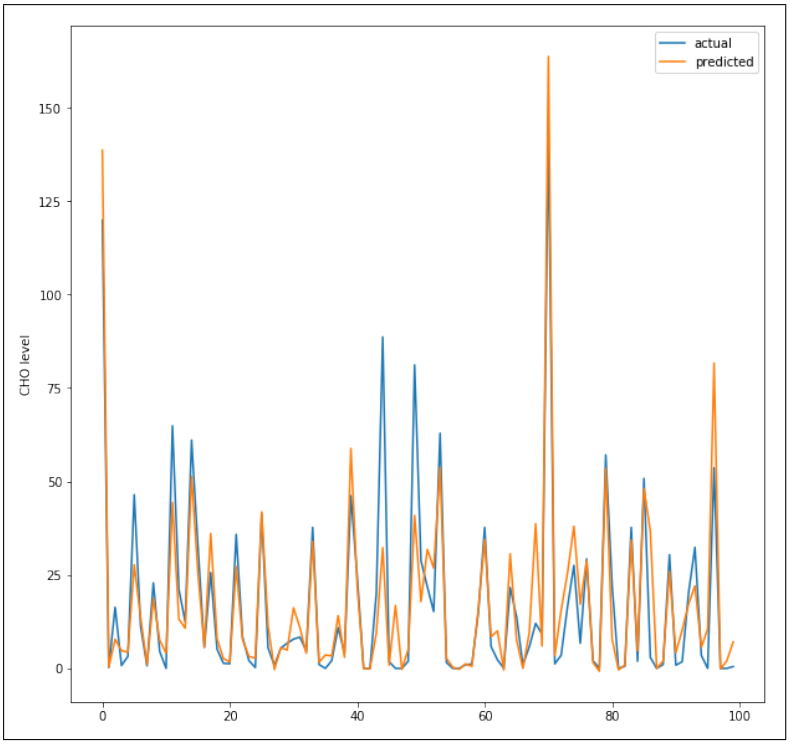

Figure 13: Graph showing actual and predicted values

5.2 Clustering and Linear Regression

- load data in pandas dataframe, both day1 and day2
- consider only time, eating occasion and CHO columns
- create 9 clusters using k-means
- Apply linear regression on each cluster
- Go back and create manual cluster according to eating occasion.
- Apply linear regression to each of the cluster

5.3 With Moving Average

- load data
- take moving average of time and CHO

- X = moving averaged time
- y = moving averaged CHO
- apply linear regression
- Got r2 score as 0.8

5.4 Future work

5.4.1 Robust one

https://uk.mathworks.com/videos/data-driven-robust-control-of-insulin-therapy-1550745172454.html

5.4.2 LSTM

https://www.frontiersin.org/articles/10.3389/fams.2017.00014/full

6 Professional Issues

7 Self Assessment

8 AI Ethical

9 How to Use my Project

System Requirements

References

[1] Ryan Andrews. All about nutrient timing: Does when you eat really matter. *Precision Nutrition*, 2019.

[2] Gregory L Austin, Lorraine G Ogden, and James O Hill. Trends in carbohydrate, fat, and protein intakes and association with energy intake in normal-weight, overweight, and obese individuals: 19712006. *The American Journal of Clinical Nutrition*, 93(4):836–843, 02 2011.

[3] Jason Brownlee. Linear regression for machine learning. *Machine Learning Mastery*, 2016.

[4] Jason Brownlee. Moving average smoothing for data preparation and time series forecasting in python. *Machine Learning Mastery*, 2016.

[5] Jason Brownlee. Overfitting and underfitting with machine learning algorithms. *Machine Learning Mastery*, 2016.

[6] Jason Brownlee. Supervised and unsupervised machine learning algorithms. *Machine Learning Mastery*, 2016.

[7] Apoorva Dave. Regression in machine learning. *Medium.com*, 2018.

[8] Hovorka Roman Elleri Daniela, Dunger David B. Closed-loop insulin delivery for treatment of type 1 diabetes. 2011.

[9] Terry G Jr Farmer. The future of open- and closed-loop insulin delivery systems. 2008.

[10] Centers for Disease Control and Prevention. Nhanes 2013-2014 dietary data. *National Center for Health Statistics*, 2018.

[11] Institute for Quality and Efficiency in Health Care. *Hyperglycemia and hypoglycemia in type 1 diabetes.* InformedHealth.org [Internet]. Cologne, Germany, 2007.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, second edition, 2009.

[13] Will Koehrsen. Random forest simple explanation. *Medium.com*, 2017.

[14] Nicola Paoletti, Kin Sum Liu, Scott A. Smolka, and Shan Lin. Data-driven robust control for type 1 diabetes under meal and exercise uncertainties. *CoRR*, abs/1707.02246, 2017.

[15] Peter Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, November 1987.

[16] Elliott Saslow. Unsupervised machine learning. *Medium.com*, 2018.

[17] NHS UK. Diabetes. 2019.

[18] USDA. Fooddata central. *USDA*, 2019.

[19] Volodya Vovk. General principles of machine learning. *Royal Holloway University of London*, 2018.

[20] Volodya Vovk. Linear regression. *Royal Holloway University of London*, 2018.

[21] Volodya Vovk Zhiyuan Luo. Unsupervised learning. *Royal Holloway University of London*, 2018.