Universidad de Monterrey

División de Ingeniería y Tecnologías
Departamento de Ingeniería

Final Project:
Predicting postprandial blood glucose level

Dr. Andres Hernandez Gutierrez
Inteligencia Artificial

Viviana Vázquez Gómez Martínez #509271
Juan Manuel Alvarez Sanchez #511385
Orlando Xavier Torres Guerra #513341
Mauricio De León Cárdenas #505597

San Pedro Garza García, N.L. a 28 de mayo de 2020

Index

Introduction 2

Objectives 3

Development with methodology 3

Discussion of results (document your experiment with images / videos) 9

Individual conclusions 11

References 12

Introduction

Diabetes is a disease that occurs when the blood glucose count in a person is too high. Approximately, just in the U.S, 30.3 million people have diabetes as of 2015 but nearly 7.2 million are undiagnosed. One reason for this might be that diabetes symptoms can be difficult to identify since these come on slowly and can easily be mistaken individually for other diseases. Not identifying the disease on time could lead to serious health problems such as an increase of probability in strokes, blindness, kidney failure, or even worse, death.

Mexico is the second country with the highest number of people with diabetes in North America after the USA, with 12.8 millions of people (IDF 2019).

The diabetes type which we are going to focus our work on is Diabetes Type 1, since our data was collected by a person with this type of diabetes.Type 1 occurs when the body doesn't produce enough of the hormone that allows cells to absorb and use glucose. This hormone is called insulin.

While a person can prevent type 2 by avoiding a sugar-rich diet and inactive lifestyle, preventing type 1 is not possible. The immune system attacks clusters of cells in the pancreas that would normally produce insulin, called islets, stopping or slowing insulin production.Without enough insulin, glucose cannot enter the cells and remains in the bloodstream.

A person with type 1 diabetes will need to take insulin for the rest of their life. Not doing so can result in ever-increasing blood sugar levels and dangerous complications. Type 1 diabetes can occur at any age, although it is more common in children and young adults. Insulin needs to be administered on every meal and the dose depends on the amount of carbohydrates the patient eats. The blood glucose level reaches a peak value 2 hours after eating and this value should be lower than 180 mg/dl for a patient with type 1 diabetes. Taking insulin keeps this peak value within the desired range.

Objectives

The objective of this work is to predict the peak value of blood glucose in mg/dl after 2 hours of eating, given an initial blood glucose value, amount of carbohydrates to eat, insulin units to take, time of measure and whether the patient exercised that day or not. This can help the treatment of diabetes to alert the patient if his/her estimate for insulin dose is accurate and will keep the peak value below 180 mg/dl or his/her particular goal.

Development with methodology

Data exploration stage:

Once the objectives of this project were identified we needed to observe our dataset, reflect upon what each feature meant and how they might have an impact in the result, this to decide whether or not to do feature design, feature scaling or feature selection. The original dataset looked like this:

Glucose (mg/dl)	**Carbs (grams)**	**Units (of insulin)**	**Exercise (1 = done, 0 = not done)**	**Time (24 hour format)**	**Outcome (glucose)**
87	60	12	1	14:57	186
219	30	10	1	20:39	194
270	35	6	1	9:02	185
109	45	9	1	1:45	69
195	15	7	1	19:40	100
154	60	9	0	10:50	170
105	26	5	0	14:22	97
76	40	8	1	9:07	102
121	35	7	1	14:12	174
164	45	9	1	14:15	196
84	45	9	1	9:14	148
117	45	10	1	14:17	150
91	45	9	1	13:44	82
58	30	0	1	11:57	91
222	0	9	1	6:54	54
100	45	6	1	16:59	213
241	30	9	1	22:02	176
70	35	0	1	11:49	172
159	0	6	1	6:30	82

77	45	9	1	8:35	123
125	45	9	1	13:54	201
76	45	9	1	9:11	123
108	45	9	1	14:25	78
134	45	9	1	14:04	168
188	60	6	1	14:30	235
53	30	0	1	8:26	147
138	60	6	1	8:06	200
97	60	12	1	20:57	174
178	35	6	1	9:13	152
87	60	12	1	13:49	67
121	45	9	1	13:47	82
92	45	9	0	14:36	142
223	45	12	0	13:32	253
148	45	9	1	14:52	152
199	45	6	1	9:17	168
114	30	0	1	18:47	197
52	20	6	1	14:57	61
93	45	9	1	10:18	58

Table 1. Glucose level dataset for multivariate linear regression

Feature descriptions:

- **Glucose:** Defines the glucose level of the user right before eating (measured in mg / dl).
- **Carbs:** Amount of carbohydrates consumed (measured in grams).
- **Units of insulin:** Amount of units of insulin which is injected before the meal (measured with BD insulin syringe).
- **Exercise:** Boolean which indicates if the user did exercise that day.
- **Time:** Time of the meal (measured in 24 hour format).
- **Outcome:** Level of glucose two hours after eating (measured in mg/dl).

These features were selected following the preprandial and postprandial glucose measurement theory, which allows patients to determine their expected glucose two hours after their meal. According to the medical theory the glucose peak takes effect approximately two hours later, but can vary from patient to patient. Following the theory accordingly it was decided to leave the amount of features as it is, with the exception of time, which in order to prevent problems while running the algorithm in Python needed to be changed to military time (without the ":”).

Data distribution

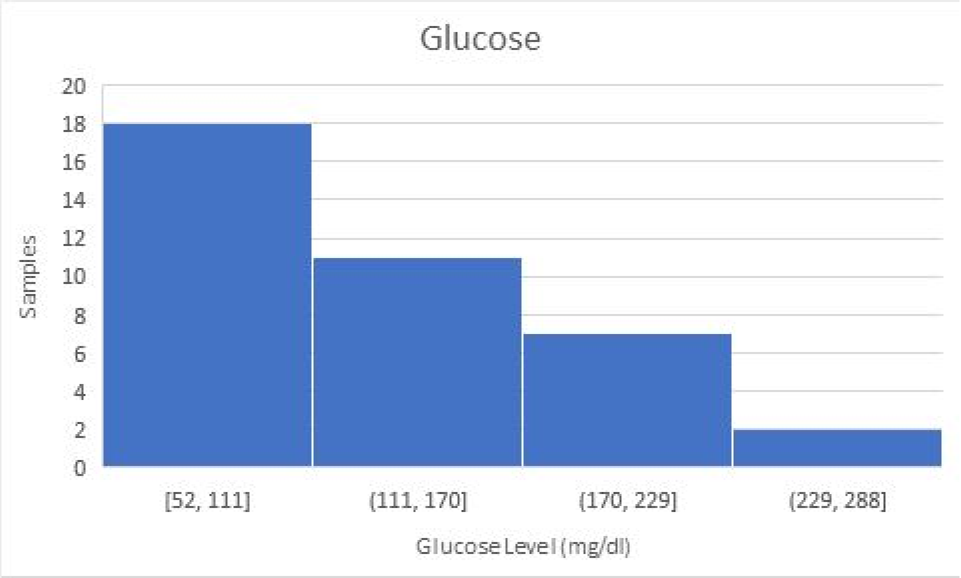


Chart 1. Glucose Level distribution

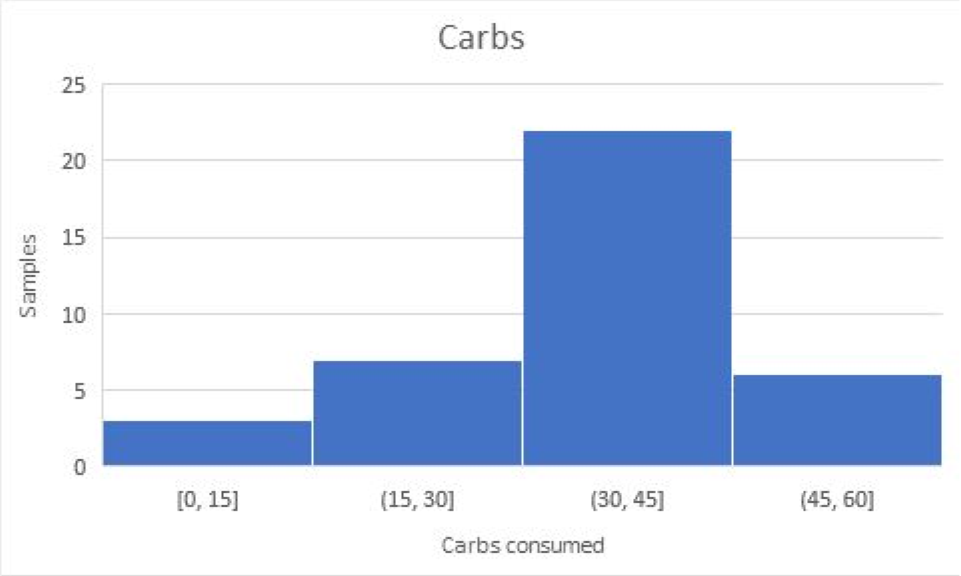


Chart 2. Carbs consumption distribution

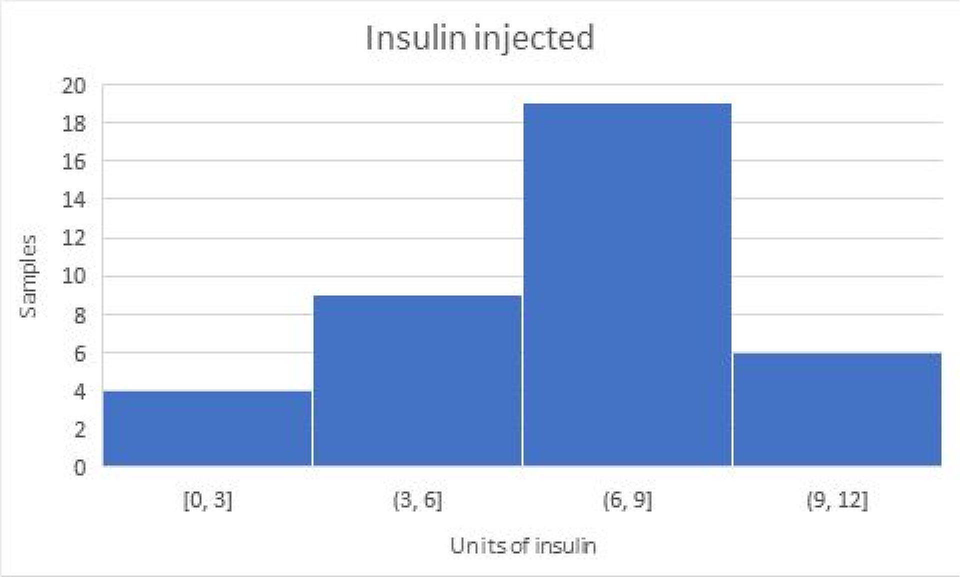


Chart 3. Units of insulin injected distribution

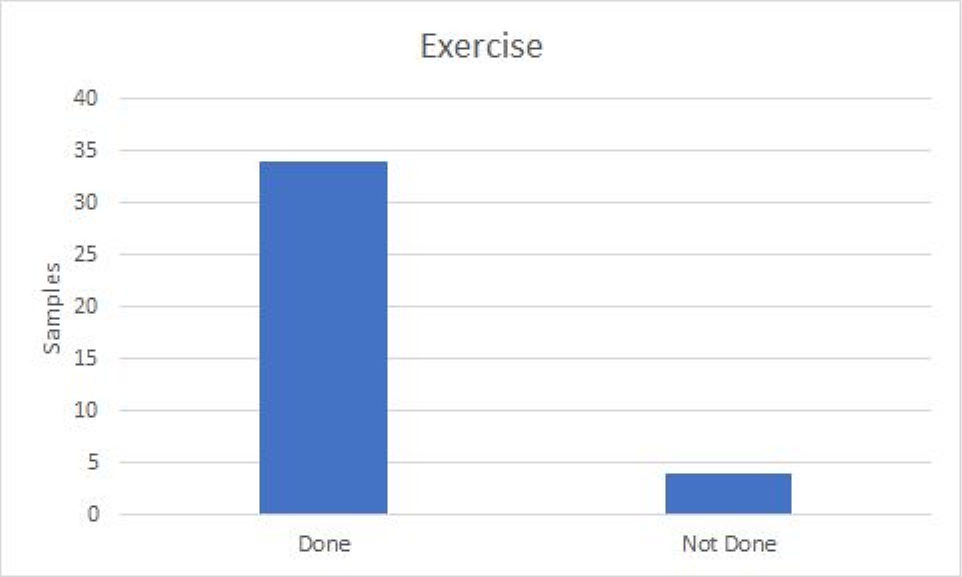


Chart 4. Exercise register distribution

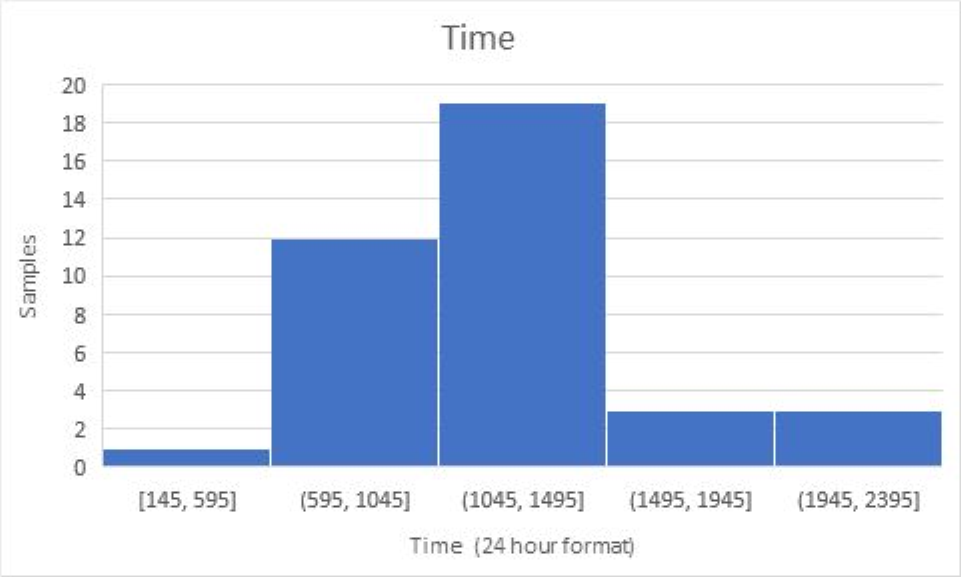


Chart 5. Time of meal consumption distribution

Once observing the outcome column in Table 1, we came to realize which machine learning algorithm was needed in order to predict those same types of results, so we opted for a multivariate linear regression algorithm, which will be explained next.

Linear regression

Linear regression is used for finding a linear relationship between target and one or more predictors. There are two types of linear regression: simple and multiple. To accomplish the objective of this work and considering that we have several features that can affect our outcome, we will use a multiple linear regression to build our model.

The multiple regression model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables.
- The independent variables are not too highly correlated with each other.
- Yi observations are selected independently and randomly from the population.
- Residuals should be normally distributed with a mean of 0 and variance σ.

When interpreting the results of a multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

Gradient Descent for multivariate linear regression

In order to obtain the best line that represents the relationship between inputs and their output, the machine learning algorithm must “train” through a mathematical process called gradient descent, which through iterations defines the beta values (w) that best represent each of the features from the dataset. This can be accomplished implementing the following formula:

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \alpha \, \nabla_{\mathbf{w}} J(\mathbf{w}) = \mathbf{w}_{j\,k} = \mathbf{w}_{j\,k-1} - \alpha \, \frac{1}{N} \sum_{i=1}^{N} \left[f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right] x_j^{(i)}$$

Figure 1. Gradient descent method equation

In a few words what the function says is: the beta (w) values for this kth iteration are equal to the beta values from one previous iteration minus the multiplication of the learning rate and the gradient of cost function.

Convergence and stopping criteria

We can say the algorithm converges (reaches its global minima) by evaluating the gradient of the cost function with the L2 norm or the euclidean norm and then comparing the obtained value with a previously defined stopping criteria, which for this project was defined as 0.01. A graphical representation of how the algorithm continues to iterate until convergence can be seen in the following image:

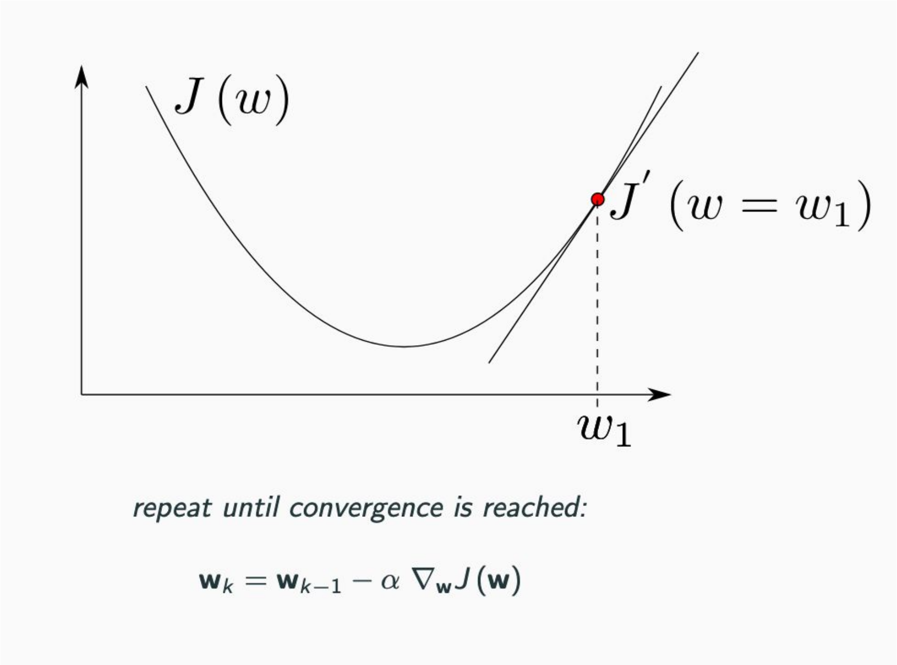

$J(w)$
$J'(w = w_1)$
w_1
repeat until convergence is reached:
$\mathbf{w}_k = \mathbf{w}_{k-1} - \alpha \, \nabla_{\mathbf{w}} J(\mathbf{w})$

Figure 2. Visual representation of the gradient descent method

Output prediction

Once the algorithm converges it’s sure to assume that the best w parameters that represent each feature were obtained, hence it’s possible to predict the expected value to new data inputs given. This can be done by multiplying the vectors with the w parameters and the vector with the input data.

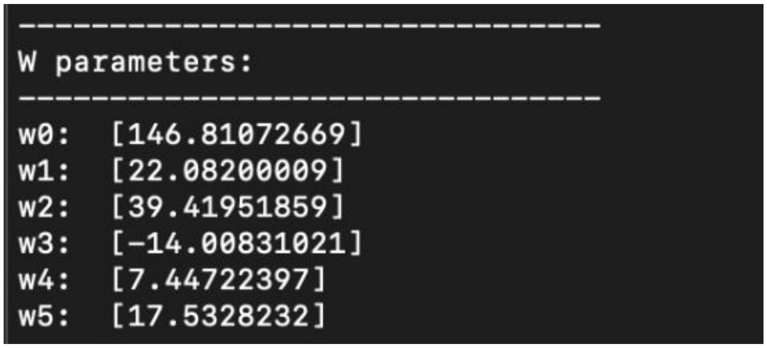
$$f_{\mathbf{w}}(x) = \mathbf{w}^T\mathbf{x}$$

Figure 3. Value prediction formula using estimated parameters

The results obtained during this project will be discussed in the following section

Discussion of results (document your experiment with images / videos)

Following the course indications and the protocol to divide the dataset for training and testing, 80% of the samples were used for training data and 20% as testing data. Given a total of 38 samples, 31 samples were used for training and 7 for testing.
The training of the model resulted in the following estimated w parameters with a learning rate of 0.0005 and a stopping criteria using the L2 norm of 0.01.



```
--------------------------------
W parameters:
--------------------------------
w0:  [146.81072669]
w1:  [22.08200009]
w2:  [39.41951859]
w3:  [-14.00831021]
w4:  [7.44722397]
w5:  [17.5328232]
```

Figure 4. Estimated w parameters for multivariate linear regression

In Figure 4, it is observed that the feature with the most weight, not considering the y intercept, is the number of insulin units taken at the meal, which corresponds to w2, followed by the grams of carbohydrates consumed. This was expected since these are the factors that contribute the most to the postprandial blood glucose value.

After calculating the w parameters for the model, a prediction with the testing data was generated. Results are presented in the following table.

Testing sample	Actual output	Predicted output
1	142	111.770
2	253	145.534
3	152	159.268
4	168	189.232
5	197	190.130
6	61	78.200
7	58	120.107

Table 1. Comparison between Actual output and Predicted output given by our algorithm

The model results in shown Table 2 are good when the initial blood glucose value is between 60 and 200 mg/dl. This may be due to the fact that there were a few data samples with these values, meaning this model is not very robust for outlying values.

Discussion

The prediction obtained with this model will be useful for a patient who is controlled most of the time, having a few outlying values for the initial blood glucose level. In sample 6 of the testing data, the model could have predicted a hypoglycemia (low blood glucose level) and the patient could have taken less insulin to avoid this condition.

Blood glucose meters have an advertised uncertainty of up to ±25mg/dl of the glucose value. Given this uncertainty, the model prediction for an initial blood glucose value between 60 and 200 mg/dl is accurate on par with market devices.

Outlying values are important since many external factors such as an illness, stress, excess of exercise, can alter the blood glucose level of the patient even if a proper control is strictly followed. However, it is a complex task to predict outlying values accurately without the proper amount of data. The collection of the data is a strict method that most patients with more than 2 years of being diagnosed hardly follow, as experience and patient’s estimates become more accurate and this data logging is skipped.

At last, this parameter estimate works only for a single patient since every patient has a different sensibility to insulin dose, carbohydrate intake, exercise, and may have different habits. Every patient will need to log his/her own data to the algorithm to produce an accurate prediction.

Future work could aim to cluster patient patterns to create accurate prediction without a vast amount of data per patient, but a vast collective amount of data with identified similarities. The future approach could be implemented with algorithms such as the K Means Algorithm.

General Conclusion

The implemented model for multivariate linear regression using the gradient descent method for predicting postprandial glucose level is useful when the initial blood glucose level is between 60 and 200mg/dl for a 21 year old male patient. This is due to the fact that most of the data samples from the dataset were in this range for initial blood glucose value.

For the most part, the algorithm is on par with the uncertainty of market devices for measuring blood glucose level which is ±25 mg/dl. Future work will need to make the algorithm robust to outlying values.

Individual conclusions

- **Viviana:**
    - This work was more difficult than I expected, to know new things about a disease that affects a lot of people and in Mexico it is one of the deadliest and most expensive diseases and there are not many innovative products or tools in this field. We know that our model has a considerable error and that can be managed by having more data and this is the first try of a bigger report and a new approach to have an algorithm that can be implemented in an application to make it easier for people in insulin calculations to reach the desired glucose after 2 hours of eating. We have a lot to do in the area and I know this work can change something.
- **Juan Manuel:**
    - Model implementation was performed well with the available data. The objective of this work was to solve a recurrent problem within Type 1 Diabetes treatment. It is a problem I live with everyday. However, this implementation will require a vast number of data samples to make the algorithm more accurate and robust. Results were good when initial glucose was between 60 and 200 mg/dl. Predicted values with the initial blood glucose in this range are on par with market devices which demonstrates a good result. However, this method may not be very practical for most patients since it requires heavy data logging that most patients do not execute. As a person that lives with Type 1 Diabetes and has participated in helping other early patients, this can be a helpful tool for their adaptation process, when the patient has less than 2 years of being diagnosed.

Personally, I am looking forward to extending this work for a real application that can help our community.

- **Orlando:**
    - I think that multilinear regression is a very useful algorithm to make predictions about some diabetes aspects, in this case, we are looking to predict the level of glucose that a person will have 2 hours after a meal taking into consideration the features mentioned before in this work. Even though we have a considerable error margin, it can be caused by the lack of data as in our project we needed to collect the data and it wasn't any database available that works for this project purpose, we didn't have enough time to collect enough data to make a more accurate model since Juan Manuel, who was who collected the data, could only take 3 sets of data per day, due to 3 meals a day.

      Anyway, I am satisfied with the results of our model and I think that the knowledge that we obtained along the semester can be reflected in this work.
- **Mauricio:**
    - Even though the predicted outputs were not as good as expected I truly believe the reason for this is the total amount of data used for training. The more data used for training the more gradient descent iterations done hence better w parameters hence better outputs. Also, as a note, if this algorithm gets to be implemented in an app which allows patients to determine their glucose level after each meal the algorithm will have to be training every time the app is used, this because even though the theory says it’s the same for everyone, most of the times the outcomes vary from person to person and the type of diabetes. I believe if this algorithm gets perfected and implemented in an app it could help a lot of people with diabetes which still have little to no clue on how to measure their glucose levels or what to expect if they eat a certain type of food (this could also be a feature to be added on future improvements).

References

- Nall, R. (2020). *An overview of diabetes types and treatments.* Retrieved on april 19th 2020 from: https://www.medicalnewstoday.com/articles/323627
- World Health Organization. (2018). *Diabetes*. Retrieved on april 26 2020 from: https://www.who.int/news-room/fact-sheets/detail/diabetes
- National Institute of Diabetes and Digestive and Kidney Diseases. (2016). *What is Diabetes?.* Retrieved on april 26 2020 from: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes
- Miller, K. (2017). *Nearly a Quarter of People with Diabetes Don’t Know They Have It.* Retrieved on april 26 2020 from: https://www.self.com/story/undiagnosed-diabetes
- International Diabetes Federation, (2019). *IDF Diabetes Atlas. 9th edition*. [Internet] Available at: https://www.diabetesatlas.org/en/resources/

- Swaminathan, S. (2018). *Linear regression - Detailed view.* Retrieved on april 26 2020 from: https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86
- Kenton, W. (2019). Multiple Linear Regression - MLR Definition. Retrieved on april 26 2020 from: https://www.investopedia.com/terms/m/mlr.asp
- Medical News Today (2018). *Type 1 diabetes.* Retrieved on april 26 2020 from: https://www.medicalnewstoday.com/articles/323729