Artificial Intelligence: Improving the Treatment of Diabetes

Shaun Bottomley, Ewa Białas, Aaron McNulty, Fahim Rafique
UB: 20002844, 19010692, 21023611, 20009956
Department of Computer Science, University of Bradford

Abstract

Artificial Intelligence is an ever advancing and expanding area of technology that is transforming all aspects life, including the healthcare system. One disease where its treatment could be completely revolutionised and improved by the implementation of AI is diabetes as it is an extremely widespread and complex illness. The aim of this report is to develop a solution that will predict the future glucose levels of a patient using the Diabetes Data Set provided by the UCI Machine Learning Repository.

Introduction

Diabetes is a lifelong condition that causes a person's blood sugar levels to be too high and if left untreated, can cause a wide range of health complications which includes death. It was estimated that 422 million people had diabetes in 2014 which was just over 1 in 9 people (WHO, 2021). According to the Centers for Disease Control and Prevention (CDC), in America alone 34.2 million people have diabetes which is just over 1 in 10 Americans (CDC, 2020).

Most diabetes can be split into two main types: Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D), which accounts for 85-90% of all cases worldwide (WHO, 2021). T1D is where the person's pancreas fails to produce enough insulin due to the loss of beta cells which is caused by an autoimmune response. The treatment for this requires consistent doses of insulin to help manage their blood sugar levels. On the other hand, T2D is caused by the body's cells becoming resistant to insulin but as time goes on, a lack of insulin may also develop. Patients with T2D are usually treated with medicine that allows the body's cells to absorb insulin more effectively. Simple changes in lifestyle such as, having a healthy diet, weight loss and avoiding tobacco use have shown to be effective in the prevention of T2D (WHO, 2021).

In both types of diabetes, consistent and strict routines, which include the scheduling of meals, the monitoring of blood sugar levels and exercising, need to be adhered to on daily basis. In addition to this, patients are all unique so a wide range of medical factors must be accounted for, and a variety of different lifestyles must be revised. Because of all this, treatment for diabetes is very complex.

Artificial intelligence (AI) is a constantly advancing area of technology and its use in the treatment of diabetes is also increasing. AI has been used to help with the monitoring of blood glucose levels and the creation of the artificial pancreas (Contreras & Vehi, 2018). Because large amounts of data are obtained from diabetes patients, AI is needed in the analyse of it to help with using the data in beneficial ways.

Background

Artificial Intelligence Techniques

Artificial Intelligence is a branch of computer science where different methods and models are used to help solve problems and make predictions with the use of massive amounts of data. AI has three main components: an agent, an environment, and a goal. The agent must be able to give outputs to the environment which are referred to as actions and also receive inputs from the environment. The agent must also have a goal that it is trying to achieve and to make it so that the goal is not limiting and that it is flexible, the agent must receive another input, reward. The agent's goal will then be to try to maximise its reward (Legg & Hutter, 2018).

Diabetic Care

Diabetic care involves the managing of many different variables in your body which means it is a very complex and tedious process and because all patients with diabetes are different, treatment and care needs to be unique to the individual. Patients need to ensure that their daily intake of glucose does not exceed dangerous levels which means carefully reading the labels of everything they eat and drink and calculating how much glucose they have already consumed. This also means that they have to carefully schedule their meals in order to not raise their blood sugar levels too high in a short period of time. Another thing that patients need to manage is their weight. The majority of patients with type 2 diabetes are obese and so lifestyle changes must be mad and exercise routines must be planned (Hillson, 2015). T1D patients and also T2D patients in some cases, need to take regular injections of insulin to regulate their blood sugar levels. People with T1D are at risk of death without consistent insulin injections which is why it is so important that meticulous care and planning is put into managing insulin and all the other variables.

AI In Diabetic Care

Diabetic care is very complex and tedious and thus requires a lot of resources and specially trained doctors. However, not all countries have the funds or the infrastructure to properly care for patients with diabetes which is where AI would come in handy. For example, in India it is estimated that around 8-10% of the population has diabetes and because India lacks the proper resources and qualified doctors to deal with all these patients, it puts a heavy strain on the healthcare system. AI could be used to help process the data of patients to take the burden away from primary health care staff (Singla, R., et al, 2019).

One way in which AI can be implemented into the care and treatment of diabetes is through the prediction of future blood glucose values which can then be used to anticipate hyperglycaemic or hypoglycaemic events (where the blood sugar levels are either too low or too high). There is also work being done in the prediction of insulin doses before meals (Singla, R., et al, 2019).

Methodologies

Expert Systems (ES) are one of the most common types of AI systems that are used in clinical routine. The ES can capture expert knowledge, facts and reasoning techniques to help professionals produce correct diagnosis and reach more accurate conclusions. The ES has three main features with the first one being the knowledge acquisition system which is used to gather and categorise knowledge. The process of imputing data can be done either in person, by the care worker or the knowledge engineer. There is also a way of applying data from databases which hold previous studies and their outcomes. The second feature is a knowledge base which is used to store knowledge about specific issues for the ES to solve. Finally, the last feature is an inference engine which is a system that controls and implements knowledge and rules in a database to mimic deduction processes. Diabetes is slowly becoming a "civilization disease" meaning more and more people are suffering because of it. Because of this, AI is essential as it helps to make tests easier and quicker to interpret before the sickness can manifest itself and make the person ill.

The ES has variables such as rule-based seasoning (RBR), case-based reasoning (CBR), and the fuzzy system. RBR is a transfer of expert and common-sense knowledge that is loaded into a computer system and represented by if-then rules. However, real-world situations are usually very fuzzy and thus RBR is very restrictive.

The second system is CBR which has four steps: retrieve, reuse, revise and retain. CBR uses previous correct diagnoses and applies the knowledge to present and similar cases. Rigla et al (2018) explain that "case studies features need to be specified to be helpful in retrieving other cases. At the same time, features have to be discriminative enough to avoid the retrieval of cases studies which could lead to wrong solutions because of being too different." The main difference between CBR and RBR is that the former doesn't need explicit domain models, it relies only on identification of new cases with significant features, which is in fact the way CBR learns. This type of CBR has been used in a mobile application, which was developed by the Imperial College in London, that lets patients monitor their glucose levels with their smartphone. A study shows that it is more beneficial and efficient than regular bolus calculators.

The fuzzy system represents computer understanding of expertise knowledge which maps numerical inputs. For example, a blood glucose in range 80-180 mg/dl; >180 mg/dl is higher than average but <80 mg/dl is lower than average. Those results are then interpreted by the Fuzzy system with other knowledge about the patient to create the most accurate diagnosis. For instance, if sugar levels are at the higher end, the patient may not be ill yet but must stay cautious of their glucose levels. Everything above 180 mg/dl is considered a high risk of developing diabetes and around 280 mg/dl is classified as a very high risk.

Our Method

The main objective of this section is the presentation of basic regression analyses. In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the ’outcome variable’) and one or more independent variables (often called ’predictors’, ’covariates’, or ’features’).The most common form of regression analysis is linear regression, in which a researcher finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.

Theory

To find the parameters of the linear equation, we need to minimize the least squares or the sum of squared errors. Of course, the linear model is not perfect, and it will not predict all the data accurately, meaning that there is a difference between the actual value and the prediction. The error is easily calculated with,

$$\epsilon_i = y_i - \beta_0 - \beta_{11}x_i$$

and assumed to be independent and $N(0,\sigma^2)$. The linear regression curve itself is given by,

$$y_i = \beta_{00} + \beta_{11}x_{i i} + \epsilon_{i i}$$

The unknown parameters β_{00}, β_{11} and σ^2 can be estimated using the *method of least squares*. Hence, find values β_{00} and β_{11} minimizing the sum of squared residuals,

$$SSres = \sum_i \epsilon_i^2 = \sum_i \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

In order to minimise *SSres* we expand the term

$$SSres = \sum_i (y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + \beta_0^2 + 2\beta_0\beta_1 x_1 + \beta_1^2 x_i^2)$$

and calculate the partial derivative with respect to β_0 and β_1 and set them to zero.

$$\frac{\partial SSres}{\partial \beta_0} = \sum_i (-2y_i + 2\beta_0 + 2\beta_1 x_i)$$
$$0 = \sum_i (-y_i + 2\beta_{00} + \beta_{11} x_i)$$
$$0 = -n\bar{y} + n\beta_{00} + n\beta_{11}\bar{x}$$
$$\beta_{00} = \bar{y} - \beta_{11}\bar{x}$$

$$\frac{\partial SSres}{\partial \beta_1} = \sum_i (-2x_i y_i + 2\beta_0 x_i + 2\beta_1 x_i^2)$$
$$0 = -\sum_i (x_i y_i) + \beta_{00} \sum_i x_i + \beta_{11} \sum_i x_i^2$$

$$0 = -\sum_i (x_i y_i) + (\bar{y} - \beta_{11}\bar{x}) \sum_i x_i + \beta_{11} \sum_i x_i^2$$

$$\beta_{11}=\frac{\sum_i x_i y_i - \bar{y}\sum_i x_i}{\sum_i x_i^2 - \bar{x}\sum_i x_i}$$

$$\beta_{11}=\frac{\sum_i x_i y_i - n\bar{x}\,\bar{y}}{\sum_i x_i^2 - n\overline{(x)}^2}$$

Using theorem of Steiner,

$$\sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\overline{(x)}^2$$

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \left(\sum_i x_i y_i\right) - n\bar{x}\,\bar{y}$$

the coefficient β_{11} is given as,

$$\beta_{11} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

and the variance is given by,

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i (y_i - \beta_{00} - \beta_{11}x_i)^2$$

Analysis

Data Processing Steps Taken

1. Using Dataset https://archive.ics.uci.edu/ml/datasets/diabetes
2. Extracted Data to Folder using: tar xvf diabetes-data.tar.z
3. Combined files data-01 -> data-10 into 1 file and converted it into a csv file, data-01-10-mod.csv.
4. Filtered out the values that have codes 48, 57, 58, 59, 60, 61, 62, 63, 64 which are all glucose measurements.
5. Rounded all time stamps to closest hour, in order to make the box plots.
6. Removed dates and codes since they are unused after filtering.

We used time as an independent variable in our linear regression AI model and glucose level as a response variable so that given any time of the day, our AI model can predict the glucose level which could be used to help with managing living with diabetes, such as when snacking is needed or when insulin is needed, depending on the type of diabetes.

Calculation

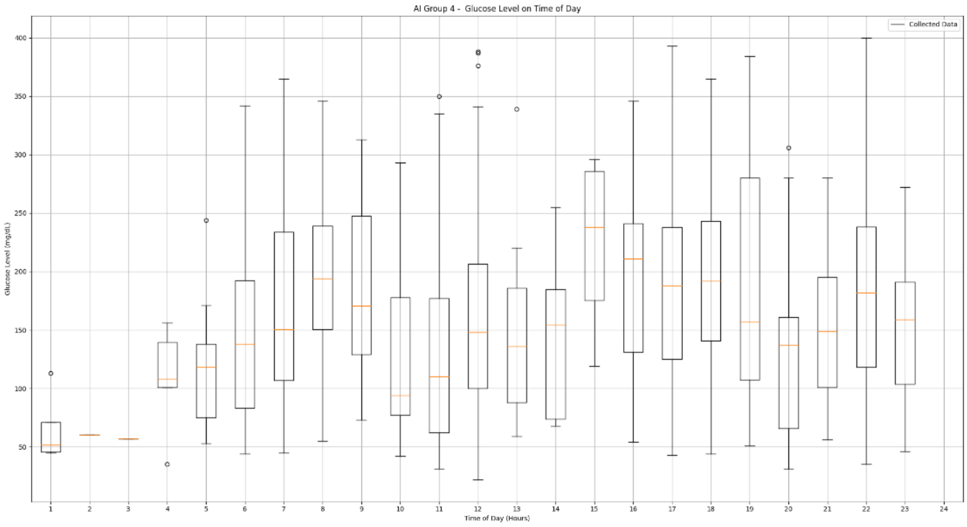
Our Regression Model gave us the following results:

Multiple R-squared	0.005249
Adjusted R-squared	0.004659
Coefficient - t value	Intercept: 38.768
	Value: 2.983
p-value	0.002899

(Done in R-Studio)

The R-squared (R^2) statistic reflects how well the model fits actual data, as it reveals how linear the relationship between our predictor (time) and response variable (glucose value) is. It always lies between 0 and 1 (i.e.: close to 0 denotes a regression model that does not adequately explain the variance in the response variable, whereas close to 1 denotes a regression that does.). For our dataset, the R^2 we get is 0.005249. So, we can say that glucose level is related to the time of the day and required measures can be taken accordingly. For example, eating food or doing exercises.

The coefficient t-value measures how far away from 0 our coefficient estimate is from 0. If it is far from zero, it would indicate that the null hypothesis cannot be rejected - thus, it is possible to declare a relationship between Time and glucose level. In our model, the t-statistic values are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists.

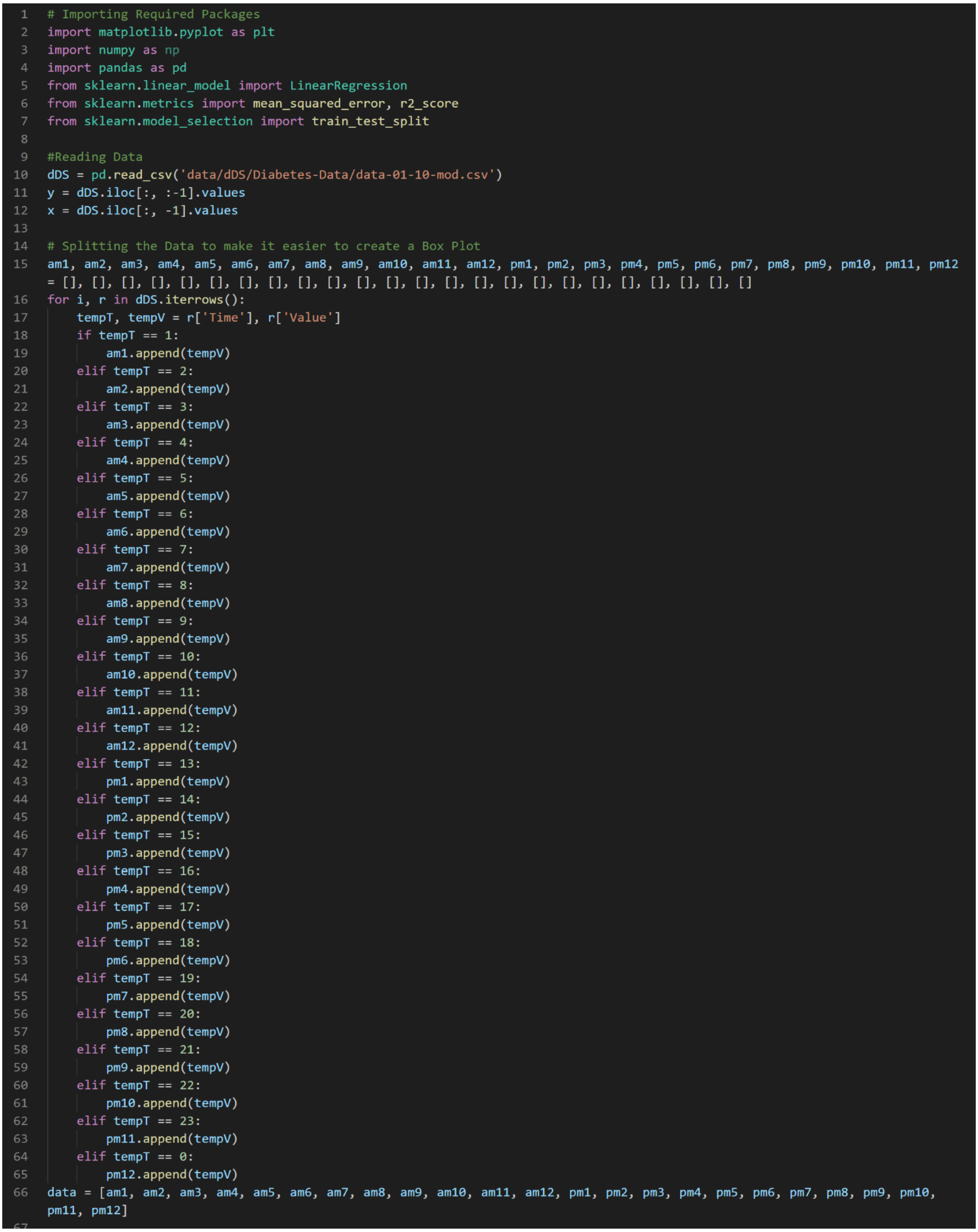


The low p-value 0.002899 tells us that there is a positive relation between the predictor and response. To predict the weight of new persons, we can use the predict() function in R.

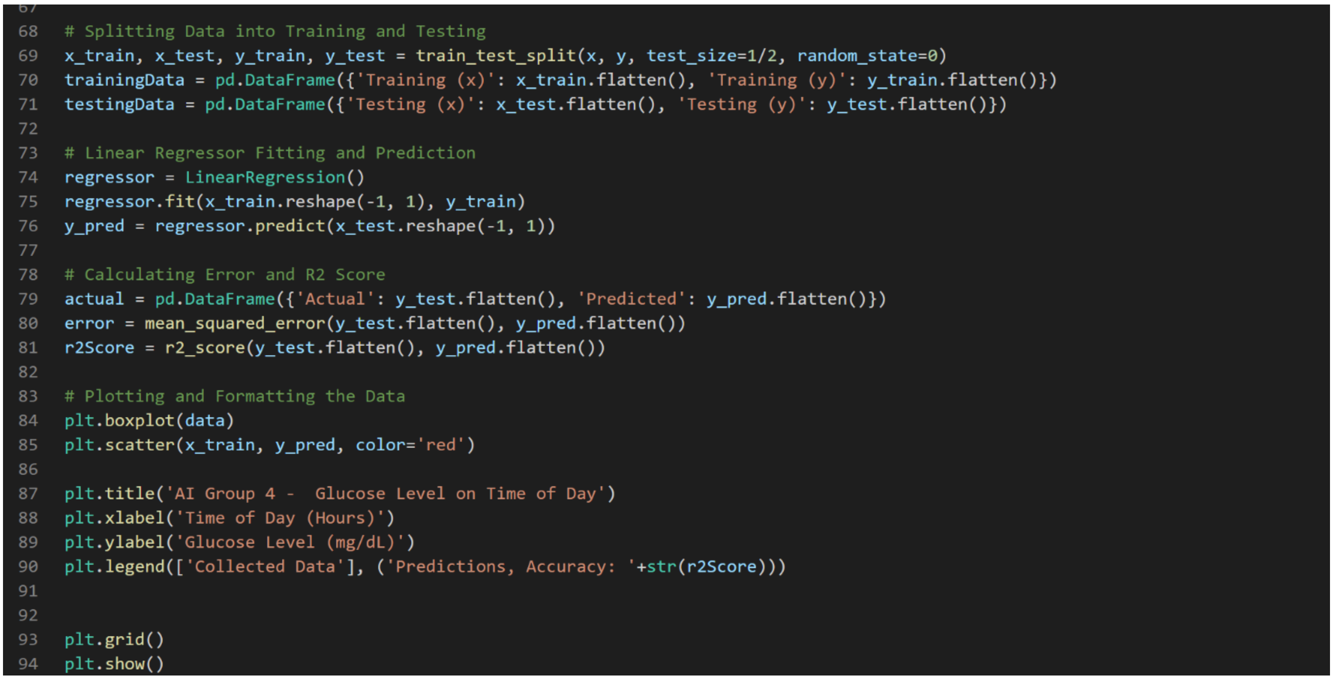
Conclusion

In conclusion, from our linear regression AI model we can say that the glucose level is positively related to the time of the day for diabetic patients. For future work, more of the data set could be used and different datasets could also be used in order to make the predictions more accurate and thus make the AI model more accurate. The use of multiple regression analysis instead of linear regression could also be implemented to further increase the accuracy of the predictions. Furthermore, other variables such as exercise activity, insulin dosage and ingestion levels could also be included in the algorithm. We can then check if these variables are related or not.

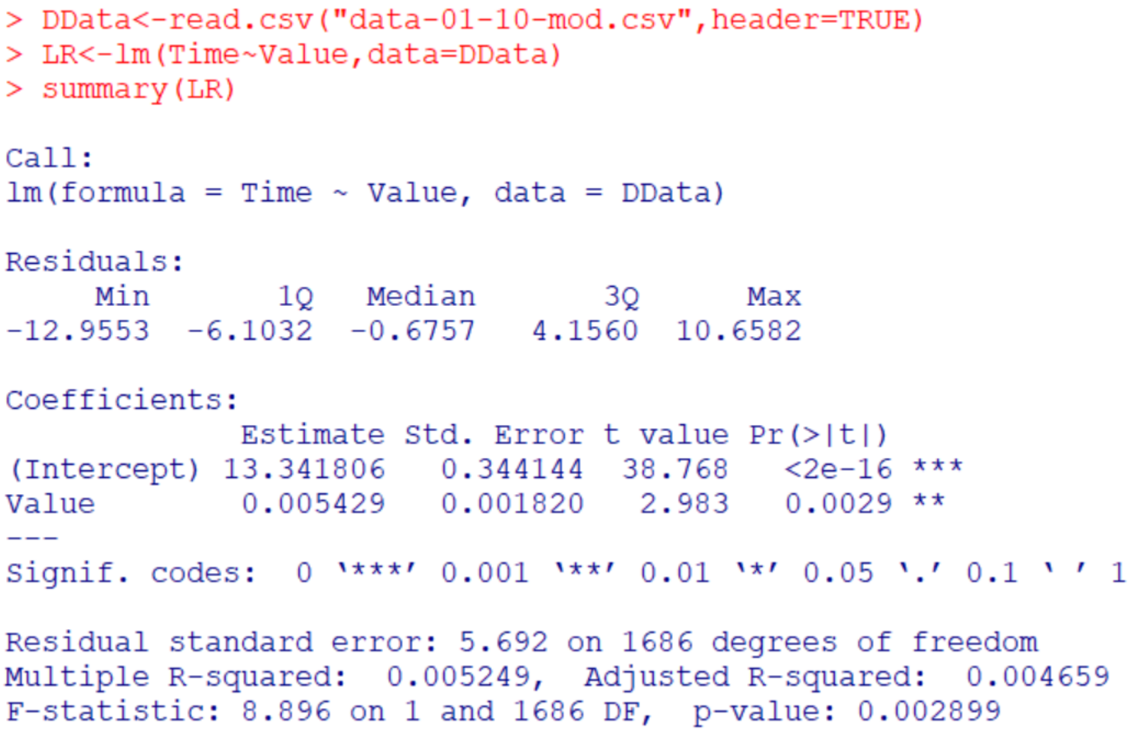
Implementation



```python
# Importing Required Packages
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split

#Reading Data
dDS = pd.read_csv('data/dDS/Diabetes-Data/data-01-10-mod.csv')
y = dDS.iloc[:, :-1].values
x = dDS.iloc[:, -1].values

# Splitting the Data to make it easier to create a Box Plot
am1, am2, am3, am4, am5, am6, am7, am8, am9, am10, am11, am12, pm1, pm2, pm3, pm4, pm5, pm6, pm7, pm8, pm9, pm10, pm11, pm12 = [], [], [], [], [], [], [], [], [], [], [], [], [], [], [], [], [], [], [], [], [], [], [], []
for i, r in dDS.iterrows():
    tempT, tempV = r['Time'], r['Value']
    if tempT == 1:
        am1.append(tempV)
    elif tempT == 2:
        am2.append(tempV)
    elif tempT == 3:
        am3.append(tempV)
    elif tempT == 4:
        am4.append(tempV)
    elif tempT == 5:
        am5.append(tempV)
    elif tempT == 6:
        am6.append(tempV)
    elif tempT == 7:
        am7.append(tempV)
    elif tempT == 8:
        am8.append(tempV)
    elif tempT == 9:
        am9.append(tempV)
    elif tempT == 10:
        am10.append(tempV)
    elif tempT == 11:
        am11.append(tempV)
    elif tempT == 12:
        am12.append(tempV)
    elif tempT == 13:
        pm1.append(tempV)
    elif tempT == 14:
        pm2.append(tempV)
    elif tempT == 15:
        pm3.append(tempV)
    elif tempT == 16:
        pm4.append(tempV)
    elif tempT == 17:
        pm5.append(tempV)
    elif tempT == 18:
        pm6.append(tempV)
    elif tempT == 19:
        pm7.append(tempV)
    elif tempT == 20:
        pm8.append(tempV)
    elif tempT == 21:
        pm9.append(tempV)
    elif tempT == 22:
        pm10.append(tempV)
    elif tempT == 23:
        pm11.append(tempV)
    elif tempT == 0:
        pm12.append(tempV)
data = [am1, am2, am3, am4, am5, am6, am7, am8, am9, am10, am11, am12, pm1, pm2, pm3, pm4, pm5, pm6, pm7, pm8, pm9, pm10, pm11, pm12]
```



```
# Splitting Data into Training and Testing
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=1/2, random_state=0)
trainingData = pd.DataFrame({'Training (x)': x_train.flatten(), 'Training (y)': y_train.flatten()})
testingData = pd.DataFrame({'Testing (x)': x_test.flatten(), 'Testing (y)': y_test.flatten()})

# Linear Regressor Fitting and Prediction
regressor = LinearRegression()
regressor.fit(x_train.reshape(-1, 1), y_train)
y_pred = regressor.predict(x_test.reshape(-1, 1))

# Calculating Error and R2 Score
actual = pd.DataFrame({'Actual': y_test.flatten(), 'Predicted': y_pred.flatten()})
error = mean_squared_error(y_test.flatten(), y_pred.flatten())
r2Score = r2_score(y_test.flatten(), y_pred.flatten())

# Plotting and Formatting the Data
plt.boxplot(data)
plt.scatter(x_train, y_pred, color='red')

plt.title('AI Group 4 -  Glucose Level on Time of Day')
plt.xlabel('Time of Day (Hours)')
plt.ylabel('Glucose Level (mg/dL)')
plt.legend(['Collected Data'], ('Predictions, Accuracy: '+str(r2Score)))


plt.grid()
plt.show()
```



```
> DData<-read.csv("data-01-10-mod.csv",header=TRUE)
> LR<-lm(Time~Value,data=DData)
> summary(LR)

Call:
lm(formula = Time ~ Value, data = DData)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.9553  -6.1032  -0.6757   4.1560  10.6582 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.341806   0.344144  38.768   <2e-16 ***
Value        0.005429   0.001820   2.983   0.0029 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.692 on 1686 degrees of freedom
Multiple R-squared:  0.005249,	Adjusted R-squared:  0.004659 
F-statistic: 8.896 on 1 and 1686 DF,  p-value: 0.002899
```

Bibliography

Centers for Disease Control and Prevention (2020) *National Diabetes Statistics Report 2020*. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf Accessed: 23 November 2021

Contreras I, Vehi J. (2018). Artificial intelligence for diabetes management and decision support: Literature review. *Journal of Medical Internet Research*, 20(5).

Hillson, R. (2015). *Diabetes care: A practical manual* (Second ed.). Oxford University Press.

Legg, S. & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines (Dordrecht)*, 17(4).

Rigla, M., García-Sáez, G., Pons, B., & Hernando, M. E. (2018). Artificial intelligence methodologies and their application to diabetes. *Journal of Diabetes Science and Technology*, 12(2)

Singla, R., Singla, A., Gupta, Y. & Kalra, S. (2019). Artificial intelligence/machine learning in diabetes care. *Indian Journal of Endocrinology and Metabolism*, 23(4).

World Health Organization (2021) *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes Accessed: 23 November 2021