Graphical user interface, application

Description automatically generated

Analysis of Diabetes

and recommendations for Predictive Models

B3T1101 – Data Engineering and Management Assignment 3

Group 7

Khoi Gia Pham 523755

Nam Hoang Giap 534598

Lars Ross 532836

John Tran 532841

Hong Minh Nguyen 532460

Background pattern

Description automatically generated with low confidence

Table of Contents

[Introduction 1](#_Toc99661392)

[Part 1: Individual and collective summaries 1](#_Toc99661393)

[Part 2: Behavior of blood glucose measurements daily 2](#_Toc99661394)

[Part 3: Behavior of blood glucose measurements throughout the whole timeframe 5](#_Toc99661395)

[Part 4: Conclusion 10](#_Toc99661396)

# Introduction

With the aim to build predictive models for Diabetes, the firm is making its decisions on which type and how the models will be built to invest resources properly. A significant issue when conducting Diabetes treatment is the patient’s constant need to monitor their blood sugar levels - the key indicator to identifying Diabetes. However, different patients exhibit different blood glucose levels and behavior. The values in the whole dataset reveal blood glucose measurement, the specific measurement code, and the period taken, whose changes will be the foundation and knowledge to indicate a Diabetes patient. As our group number is 7, the data we worked with belongs to patients with IDs ranging from 31 to 35. Since the data has not been ready for fast development, there will be several steps to prepare the data consisting of values representing Blood Glucose and Insulin Intake measurements for further analysis. This report will clarify our selection of the optimal way for building predictive models for Diabetes, whether it should be generic for everyone or individual models for each patient.

# Part 1: Individual and collective summaries

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient ID** | **Blood glucose levels mmol/L** | | | |
| Min | Max | Average | Standard deviation |
| 31 | 3 | 335 | 78.55 | 76.31 |
| 32 | 0 | 361 | 64.59 | 90.45 |
| 33 | 0 | 349 | 64.21 | 85.60 |
| 34 | 0 | 302 | 62.27 | 80.84 |
| 35 | 0 | 302 | 59.28 | 79.09 |
| **All patients** | **0** | **361** | **68.61** | **80.91** |

Table 1. Summaries of Blood glucose levels

As shown in Table 1, there is a significant difference between each patient’s minimum and maximum blood sugar values. The standard deviations are relatively high compared to the average values, implying a positively skewed distribution and high variations in the patients’ blood sugar values. Thus, patients’ blood sugar values are not similar in the long term.

Further, the “All patients” row indicates the collective data of all five patients’ blood sugar values. The minimum value is 0, the same as those of patients 32 to 35. The maximum value is 361, the same as for patient 32, implying that patient 32 has reached the highest blood sugar value out of the five patients. Therefore, patient 32 has the broadest range of blood sugar values. Patient 32 also has the highest standard deviation of 90.45. Interestingly, Patient 31 has the highest average value while at the same time the lowest standard deviation value compared to those of other patients. The above information reveals that patient 31 has the slightest variation in his recorded blood sugar values, which are closer to their respective mean values.

Overall, the average Blood glucose levels of Diabetes patients in the given data sets are 68.61 mmol/L. These values below 70 mmol/L thus, indicate that those patients are suffering from hypoglycemia - a life-threatening condition.

# Part 2: Behavior of blood glucose measurements daily

In this section, a validation study on the data regarding the daily behavior of blood glucose measurements is conducted. We will only consider the value corresponding to blood glucose measurements (48 to 60). To begin with, we mutate one column with a "datetime" value (using paste followed by as.POSIXct and format functions). Then we extracted the hour value from "datetime" with the help of mutate, POSIXct, and format funcitons. After that, we changed the hour into numeric so as to classify each measurement according to the time they were taken during the day. Finally, for each Hour timestamp, we calculate the mean value respectively. Figure 1 below illustrates the behavior of blood glucose measurements throughout the day.

QChart, radar chart, line chart

Description automatically generated

Figure 1. Individual blood glucose levels variation during a day of Patient ID 31 to 35

In detail, the first patient, 31, is measured with paper records. These records were taken at fictitious times, 08:00, 12:00, 18:00, and 22:00. The other four are measured with an automatic electronic recording device. Consequently, more real-time stamps were recorded. When comparing 31 with the others, the points at the given time stamps seem relatively close. Therefore, we doubted that the graph could have looked similar if patient 31 had been given an automatic device.

The other patients, 32, 33, 34, 35, all have a decrease in glucose levels overnight, which is normal. However, the decreasing amount proves problematic. Patient 35 has a deficient level in the morning, which indicates hypoglycemia. The others start their day at a normal level, except patient 31. This patient data was obtained from paper records, and the data have fictitious uniform recording time. Thus, the average value by hour remained nearly unchanged at around 155. Next, the normal rises**in blood glucose levels after meals are also observed. S**pecifically, there were surges in the patient’s blood glucose levels at around 6:00, 13:00, and 18:00. The figures usually rose to between 130 and 150 mmol/L during these periods.

However, patients 32 and 35 have the most significant swings in their glucose levels. They both peak far above 200 mmol/L, which usually indicates a medical condition. Both of these graphs have very deep lows as well, just like patient 33. Moreover, as discussed before, this indicates hypoglycemia, a condition applicable to these three patients. Patient 34 seems to have a normal health condition; the graphs go over thresholds for a normal person. The glucose levels drop beneath 100 mmol/L. Patient 31 looks very steady, but we miss real-time information.

Lastly, after a meal, the glucose levels should rise, and within 2 hours, they should be normal again. For all patients except patient 31, we assumed they had eaten whenever their graph rose. Therefore, the glucose levels should be normal in 2 hours. The graph shows that this is not the case for the other four patients. When they eat, the levels can rise far above normal. Additionally, the time frame is way longer than 2 hours. For example, the graph of patient 35 started rising around 04:00 and had a “normal” glucose level around 11:00. In this example, a very steep decrease happened, which is another problem for the patients. They fluctuate very highly, after a high peak comes a shallow valley.

Next, we want to examine whether collective data represent the individual patient. We merge all the patient data again and calculate the average measurement value at a particular timestamp. The collective blood glucose levels’ daily variation is shown in Figure 3.

Chart, line chart

Description automatically generated

Figure 3. Collective blood glucose level variation during a day of patient ID 31 to 35

Overall the collective line graphs bear some similarities with the line graphs of the individual patient, except for patient 31, to a certain extent. We believe that patient 31’s strange, normalized glucose levels trendlines are due to paper records. So far, we conclude that collective data represent patients whose data is obtained from the automatic electronic recording device but not from paper recording.

However, the following adverse characteristics of the collective graph should be noted. We can observe that significant drops (below 100mmol/L) often occur, and the duration is too long. However, the differences are more decisive. Moreover, in the collective graph, patients tend to go over the 200 mmol/L and under the 70 mmol/L at 9:00 and 22:00, respectively. The events indicate hypoglycemia and medical conditions, but they are not the case for all patients.

# Part 3: Behavior of blood glucose measurements throughout the whole timeframe recorded

In the first place, we tidied our data by categorizing blood glucose measurement codes as “1” and insulin intake measurement codes as “0”. Next, we eliminate all codes indicating medical issues as well as duplicate values in the combination of DateTime and Code. Afterward, a similar method from Part 2 was applied. We derived the mean value of measurements in each date group. The graphs for the patterns and further details about the Diabetes situation of each patient are presented as follows

**Chart, line chart, histogram

Description automatically generated**

Figure 4. Patient 31 - Individual Blood Glucose Levels Variation Over Dates

Above is the graph showing the changes in blood glucose measurements of patient 31, which was recorded during the period from May 1991 to mid-August 1991. Compared to other patients, the measurements of patient 31 were taken in 1991, which was far later than the others, whose recorded year was 1990. It can be seen that there have been significant changes in the blood glucose measurements of patient 31 day by day. The first measurement was around 99 mmol/L and started fluctuating from 70 to approximately 200 until the 10th of July 1991, indicating that the patient was not in a life-threatening medical condition.

Afterward, the range of fluctuation had been increased to more than 250 mmol/L, and it showed more times of passing the value barrier of 200 mmol/L, showing the medical condition of patient 31 during the time. After experiencing a decrease by approximately 125 mmol/L to 100 mmol/L, a regular measurement, patient 31’s measurements had increased significantly and peaked at around 335 mmol/L and showed no signs of decreasing until the end of the recorded time.

Chart, line chart

Description automatically generated

Figure 5. Patient 32 - Individual Blood Glucose Levels Variation Over Dates

Starting at approximately 240 mmol/L, patient 32 was in a medical condition at the beginning of the research period starting from the final week of July 1990 to the first week of August 1990. Afterward, the blood glucose measurement of patient 32 had witnessed a dramatic decline and reached the bottom of the whole research measurements records at 80 mmol/L on the very the 1st of August 1990. The measurements continued to go upwards, increasing to 200 mmol/L before the 6th of August 1990 and declined gradually afterward until the end of the research period, 8th of August 1990. The measurements showed that patient 32 was not having any of the conditions concerning medical and life-threatening possibilities most of the time.

Chart, line chart

Description automatically generated

Figure 6. Patient 33 - Individual Blood Glucose Levels Variation Over Dates

Similar to patient 32, the records of patient 33 started at the very high point of blood glucose measurements with 320 mmol/L on the ending day of July 1990. Later, there was a plummet in the values one day after the beginning of the research period. The fluctuation pattern of patient 33’s records is considered quite significant and reached 250 mmol/L, which was the second highest throughout the period, indicating the medical condition of patient 33 was going on. At the end of the research, 29th of August 1990, the blood glucose measurement of patient 33 was recorded as 120 mmol/L, revealing the situation of having no harmful conditions in the patient.

Chart, line chart

Description automatically generated

Figure 7. Patient 34 - Individual Blood Glucose Levels Variation Over Dates

When it comes to patient 34, the starting and ending points of the records did not have a significant gap in terms of value, which was different from all three patients above. The recorded value began at 139.5 mmol/L on the 23rd of August 1990 and ended at 121.75 mmol/L on the 21st of September 1990. As can be seen, the pattern of fluctuation in patient 34’s blood glucose measurements occurred every approximately equal period. Starting at the low value after the first decline, values started to go up and peaked at around 235 mmol/L on the 27th of August 1990, indicating the medical condition of patient 34. Afterward, the patient saw a significant decrease after peaking a few days and continued to increase later on, which generated a similar fluctuated pattern in the patient’s value records. Overall, for most of the days, patient 34 did not have any conditions concerning medical and hypoglycemia situations.

Chart, line chart

Description automatically generated

Figure 8. Patient 35 - Individual Blood Glucose Levels Variation Over Dates

Along those lines, patient 35’s blood glucose measurements started with a slight increase before dropping to around 126 mmol/L. The range of fluctuation for patient 35 was significant, with the span of 100 mmol/L for changes until the 1st of October 1990. During the period, the graph peaked at 198 mmol/L. A few days after the beginning of October 1990, patient 35’s measurement value reached the bottom at around 69 mmol/L, indicating that patient 35 was going through a life-threatening condition with the blood glucose measurement going below 70 mmol/L. The situation was quickly changed with a significant increase in values, and patient 35 ended the period with approximately 150 mmol/L on 10th of October 1990. Overall, compared to other patients, patient 35 was the only one who did not have a blood glucose measurement exceeding 200 mmol/L and the only one who went through hypoglycemia - a life-threatening condition.

Chart

Description automatically generated

Figure 9. Blood Glucose Change Graph for Collective Data over Dates

From figure 9, we can observe that patient 31 was measured in another year compared to the other patients. Throughout the timeframe, the blood glucose level for patients 32,33,34,35 (which were obtained from electronic recording) witnessed a slight downward trend. Meanwhile, an upward trend was witnessed in the figure for patient 31(which was obtained through paper recording). Consequently, we can conclude that patient 32, 33, 34, 35 suffers from bodily function distortion due to low blood sugar level. Patient 31, in contrast, had diabetes due to the overall level of blood sugar increasing over long periods of time

# Part 4: Conclusion

In conclusion, the findings from parts 2 and 3 above carry out similarities in terms of outcomes based on different scales of analysis. In the analysis of the variation of blood glucose measurement within a day on five chosen patients, the graphic outcome of collective data from all patients shows several similarities in patterns, compared to graphs of individual data, despite the slight difference in inpatient 31 to others. Meanwhile, concerning the changes in patients’ blood glucose measurements throughout the whole period, the graphical outcome shows clear differences. In 1991, patient 31 was diagnosed with diabetes due to high blood sugar levels. Meanwhile, in 1990, the others suffered from decreasing blood sugar levels. Moreover, fictitious uniform recording times resulting from paper records show difficulties in tracking daily blood sugar levels (reflected in patient 31 data). Hence, developing generic models for patients with electronic records proves to be prospective for the Machine Learning team. This ends the report.