A Data Driven Approach to Blood Glucose Prediction in Insulin-Dependent Diabetes Patients

Ashwini Bhatte

22/08/2017

Introduction

Insulin dependent diabetes mellitus (IDDM) also known as Type 1 diabetes is a chronic illness characterised by the body's inability to produce insulin due to the autoimmune destruction of the beta cells in the pancreas.

Patients with type 1 diabetes mellitus (DM) require lifelong insulin therapy. Most require 2 or more injections of insulin daily, with doses adjusted on the basis of self-monitoring of blood glucose levels. Long-term management requires a multidisciplinary approach that includes physicians, nurses, dieticians, and selected specialists.

Outpatient management of IDDM relies principally on three interventions: diet, exercise and exogenous insulin. Proper treatment requires careful consideration of all three interventions.

Patients with diabetes face a lifelong challenge to achieve and maintain blood glucose levels as close to the normal range as possible.

Hyperglycemia: BG $>$ 200 mg/dl

Average BG: 150 mg/dl

Hypoglycemia: BG $<$ 80 mg/dl

But it is desirable to keep 90% of all BG measurements $<$ 200 mg/dl.

With appropriate glycemic control, the risk of both microvascular and neuropathic complications is decreased markedly. In addition to being a lifelong disease, the treatment of DM is multifaceted making it necessary to take many variables into account.

Domain knowledge

Outpatient management of IDDM relies principally on three interventions: diet, exercise and exogenous insulin. Proper treatment requires careful consideration of all three interventions.

- **INSULIN**

One of insulin's principal effects is to increase the uptake of glucose in many of the tissues (e.g. in adipose/fat tissue) and thereby reduce the concentration glucose in blood.
Patients with IDDM administer insulin to themselves by subcutaneous injection. Insulin doses are given one or more times a day, typically before meals and sometimes also at bedtime. Many insulin regimens are devised to have the peak insulin action coincide with the peak rise in BG during meals. In order to achieve this, a combination of several preparations of insulin may be administered. Each insulin formulation has its own characteristic time of onset of effect, time of peak action and effective duration. These times can be significantly affected by many factors such as the site of injection (e.g. much more rapid absorption in the abdomen than in the thigh) or whether the insulin is a human insulin or an animal extract. The times I have listed below are rough approximations and I am sure that I could find an endocrinologist with different estimates.

Type of insulin	Onset of effect	Time of peak action	Effective duration
Regular Insulin	15-45 minutes	1-3 hours	4-6 hours
NPH Insulin	1-3 hours	4-6 hours	10-14 hours
Ultralente	2-5 hours	not much of a peak	24-30 hours

- **EXERCISE** Exercise appears to have multiple effects on BG control. Two important effects are: increased caloric expenditure and a possibly independent increase in the sensitivity of tissues to insulin action.

BG can fall during exercise but also quite a few hours afterwards. For instance, strenuous exercise in the mid-afternoon can be associated with low BG after dinner. Also, too strenuous exercise with associated mild dehydration can lead to a transient increase in BG.

- **DIET** Another vast subject but (suffice it to say for the purposes of users of the data set) in brief: a larger meal will lead to a longer and possibly higher elevation of blood glucose. The actual effect depends on a host of variables, notably the kind of food ingested.

For instance, fat causes delayed emptying of the stomach and therefore a slower rise in BG than a starchy meal without fat. Missing a meal or eating a meal of smaller than usual size will put the patient at risk for low BG in the hours that follow the meal.

- **GLUCOSE CONCENTRATIONS** BG concentration will vary even in individuals with normal pancreatic hormonal function.

A normal pre-meal BG ranges approximately 80-120 mg/dl.

A normal post-meal BG ranges 80-140 mg/dl.

The target range for an individual with diabetes mellitus is very controversial. I will cut the Gordian knot on this issue by noting that it would be very desirable to keep 90% of all BG measurements < 200 mg/dl and that the average BG should be 150 mg/dl or less.

Note that it takes a lot of work, attention and (painful) BG checks to reach this target range. Conversely, an average BG > 200 (over several years) is associated with a poor long-term outcome. That is, the risk of vascular complications of the high BG is significantly elevated.

Hypoglycemic (low BG) symptoms fall into two classes. Between 40-80 mg/dl, the patient feels the effect off the adrenal hormone epinephrine as the BG regulation systems attempt to reverse the low BG.

These so-called adrenergic symptoms (headache, abdominal pain, sweating) are useful, if unpleasant, cues to the patient that their BG is falling dangerously.

Below 40 mg/dl, the patient's brain is inadequately supplied with glucose and the symptoms become those of poor brain function (neuroglycopenic symptoms).

These include: lethargy, weakness, disorientation, seizures and passing out.

Objective

The goal of this project is to,

- Extract understandable patterns and associations from data through data visualisation.
- Predict the blood glucose concentration or Hyperglycemia (high blood glucose) and Hypoglycemia (low blood glucose) in patients.

Through finding a pattern in blood glucose levels and the type of activities before or after the BG measurement, we can give recommendations to the patient.

Description of Dataset

The diabetes dataset was obtained from **UCI Machine Learning Repository**, which was donated by Michael Kahn, MD, PhD, Washington University, St. Louis, MO

The dataset consists og 70 .tsv files and a data code and a domain description file. Each file out of the 70 different files represent the data for a single DM patient. Each file contain 4 attributes: Date, Time, Code, Value.

Diabetes patient records were obtained from two sources: an automatic electronic recording device and paper records. The automatic device had an internal clock to timestamp events, whereas the paper records only provided "logical time" slots (breakfast, lunch, dinner, bedtime). For paper records, fixed times were assigned to breakfast (08:00), lunch (12:00), dinner (18:00), and bedtime (22:00). Thus paper records have fictitious uniform recording times whereas electronic records have more realistic time stamps.

The Code field is deciphered as follows:

Code	explanation
33	Regular insulin dose
34	NPH insulin dose
35	UltraLente insulin dose
48	Unspecified blood glucose measurement
57	Unspecified blood glucose measurement
58	Pre-breakfast blood glucose measurement
59	Post-breakfast blood glucose measurement
60	Pre-lunch blood glucose measurement
61	Post-lunch blood glucose measurement
62	Pre-supper blood glucose measurement
63	Post-supper blood glucose measurement
64	Pre-snack blood glucose measurement
65	Hypoglycemic symptoms
66	Typical meal ingestion
67	More-than-usual meal ingestion
68	Less-than-usual meal ingestion
69	Typical exercise activity
70	More-than-usual exercise activity
71	Less-than-usual exercise activity
72	Unspecified special event

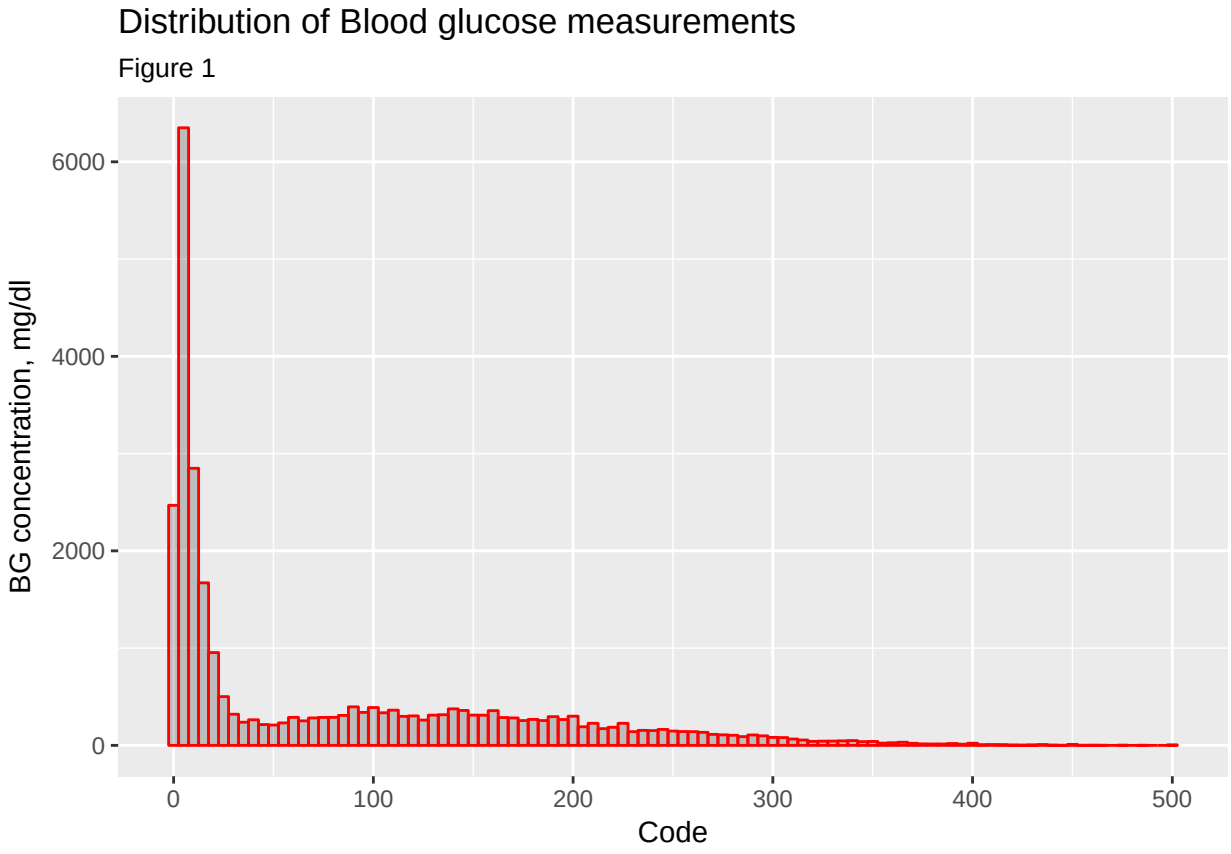
The following description of the analysis is given without the complete R code, which can be found on my **GitHub repository**.

Exploratory Data analysis

The following variables will be investigated in this section:

- Code vs Blood Glucose concentration
- Time vs Blood Glucose concentration

Since bg_conc is the only numeric variable in the data set; it???s distribution is observed by plotting the histogram.



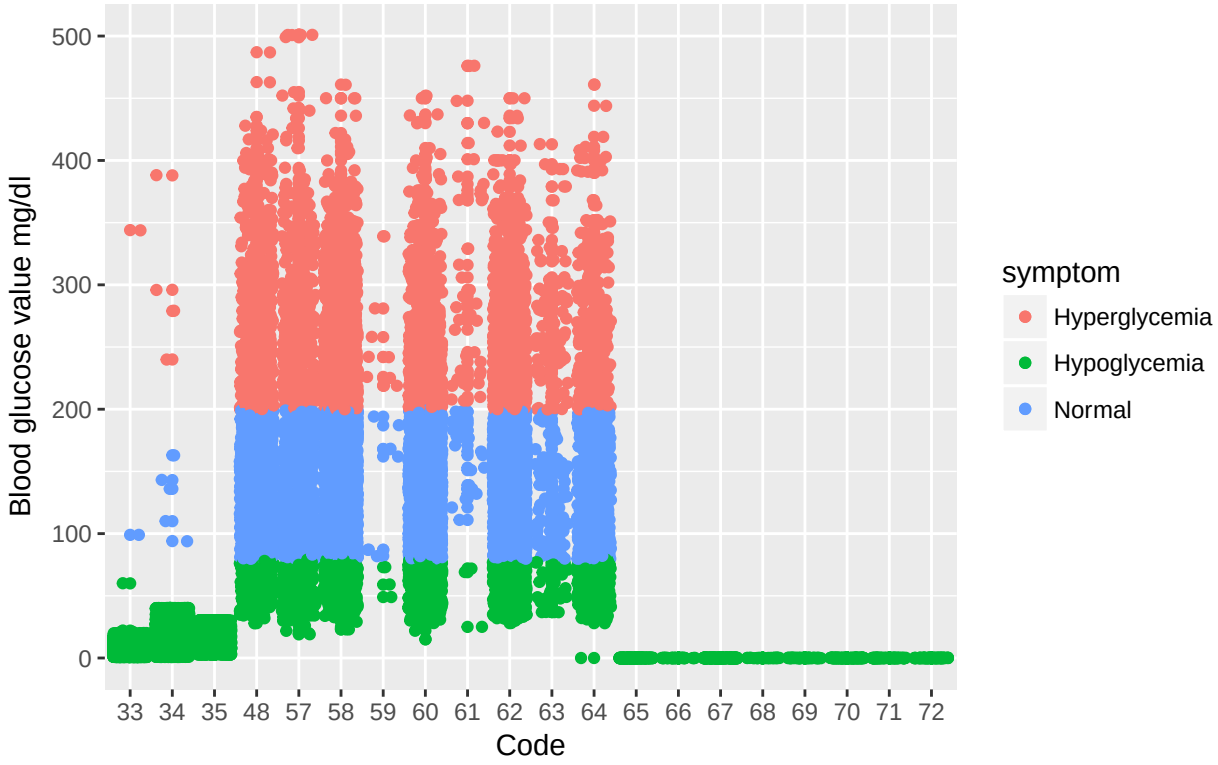
We can see the high peak at 10 - 20 mg/dl BG

Now that we have determined the distribution of BG concentration, the focus of the further data visualisation and exploration will be to deep dive into other variables and determine how they interact with one another.

Code vs Blood Glucose concentration

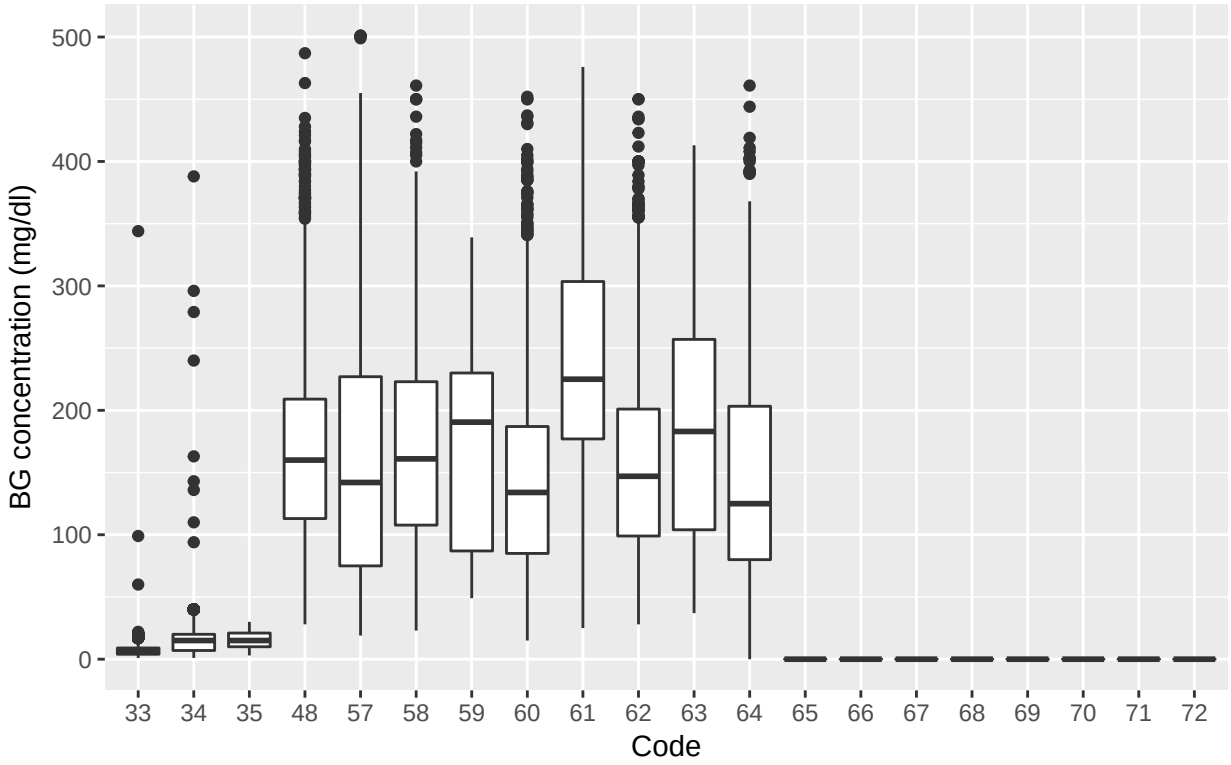
Distribution of BG measurements based on Code and BG concentration

Figure 2



Relation between Code and Blood Glucose concentration

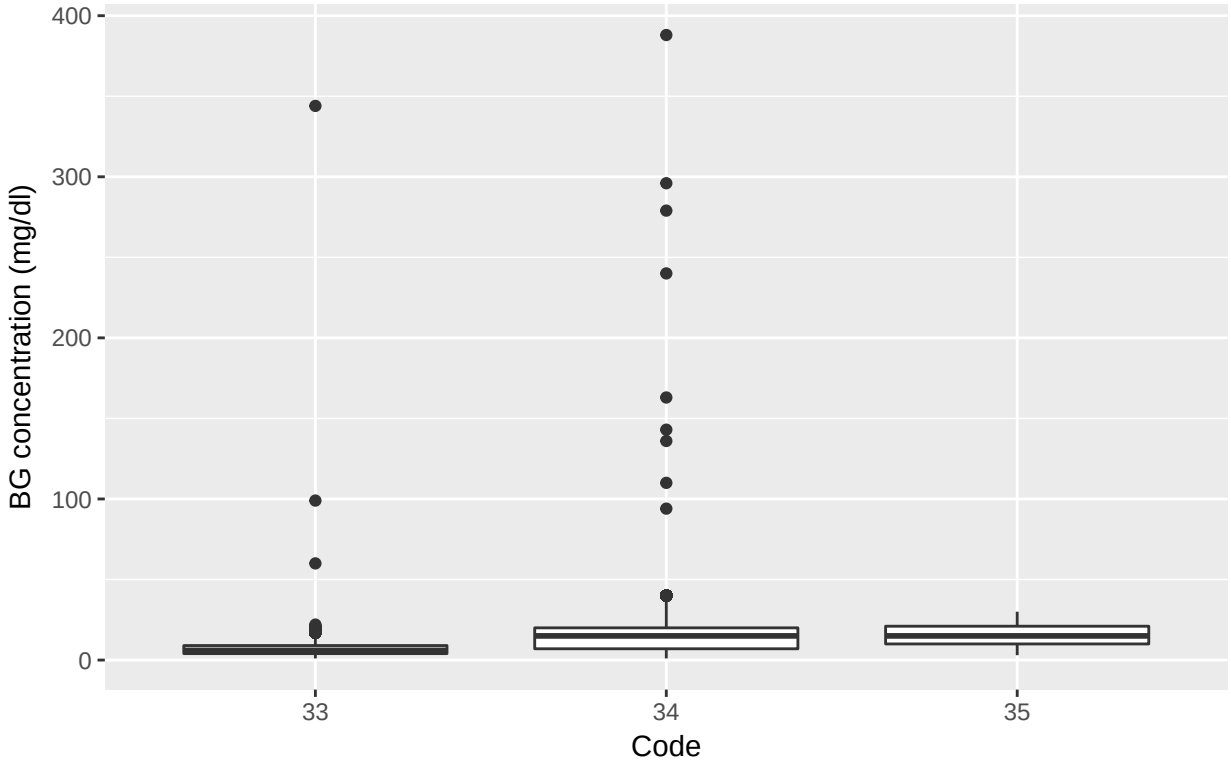
Figure 3



To gain a better perspective at this, let???s plot the graph of BG concentration vs code by grouping the codes into 3 sub categories.

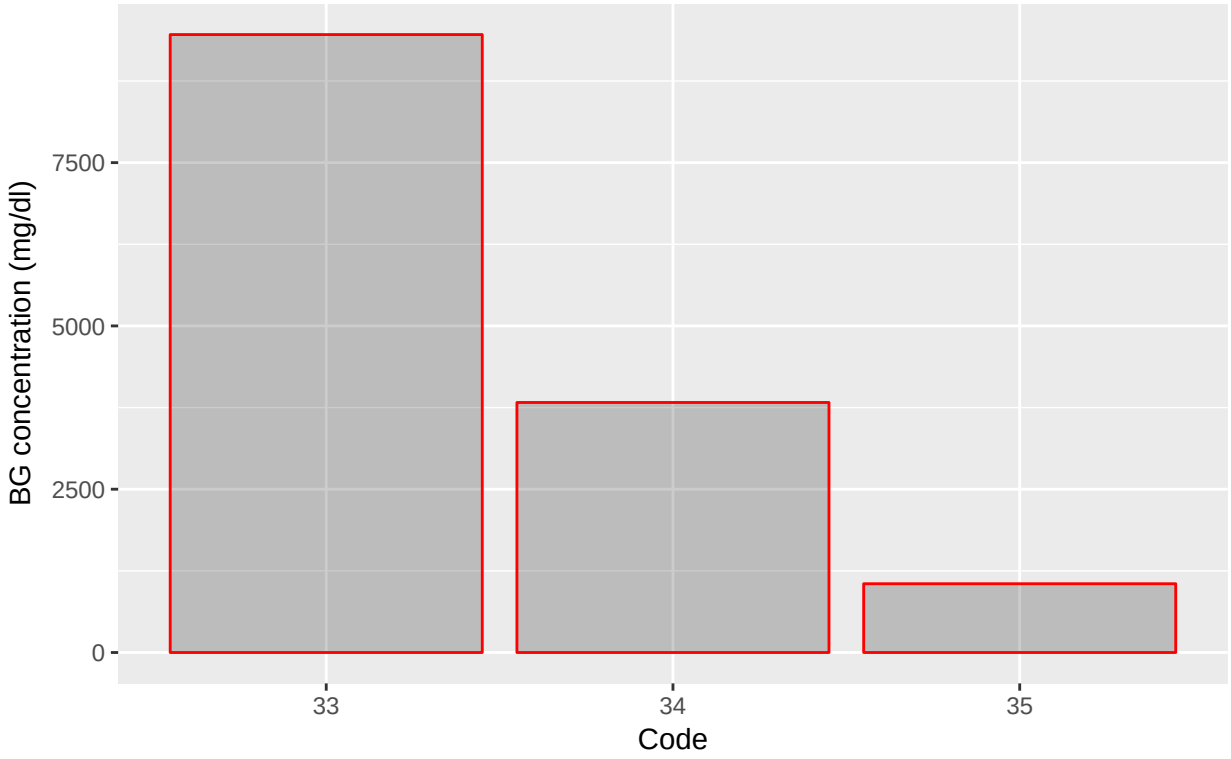
Relation between Insulin dose and Blood Glucose concentration

Figure 4



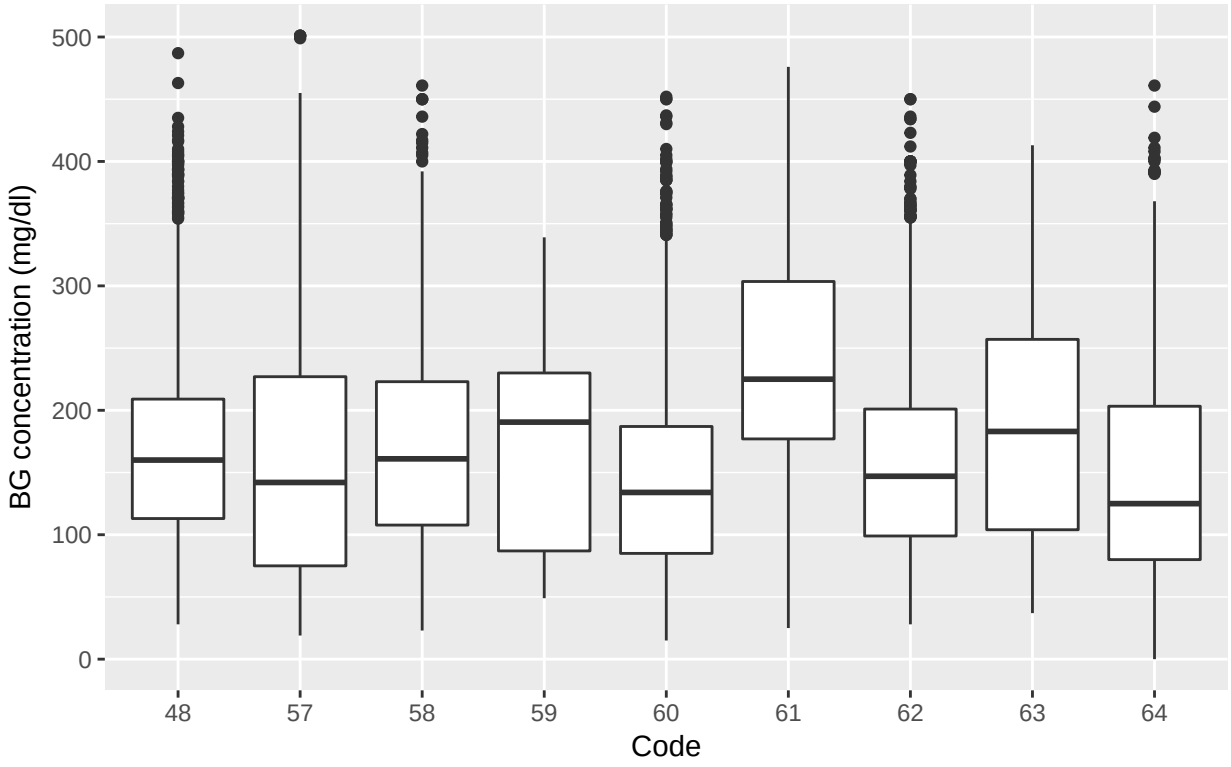
Distribution of code (Insulin)

Figure 5



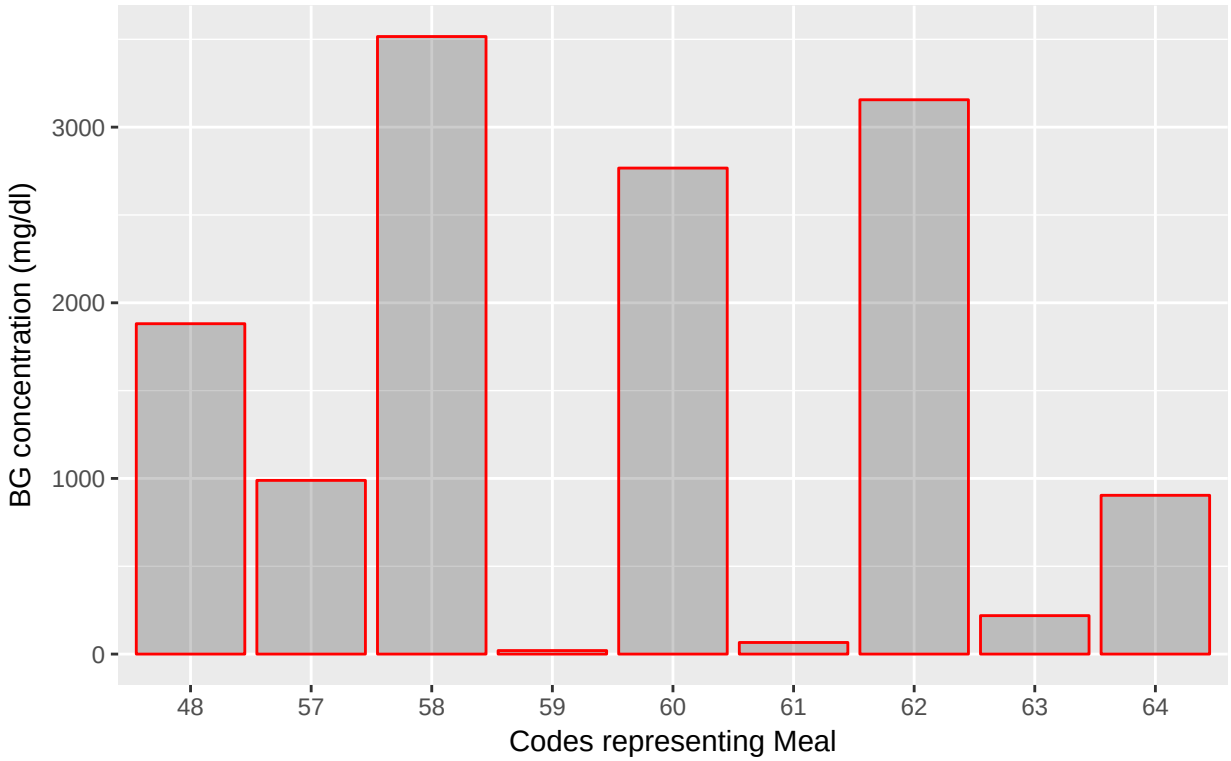
Relation between Meal and Blood Glucose concentration

Figure 6



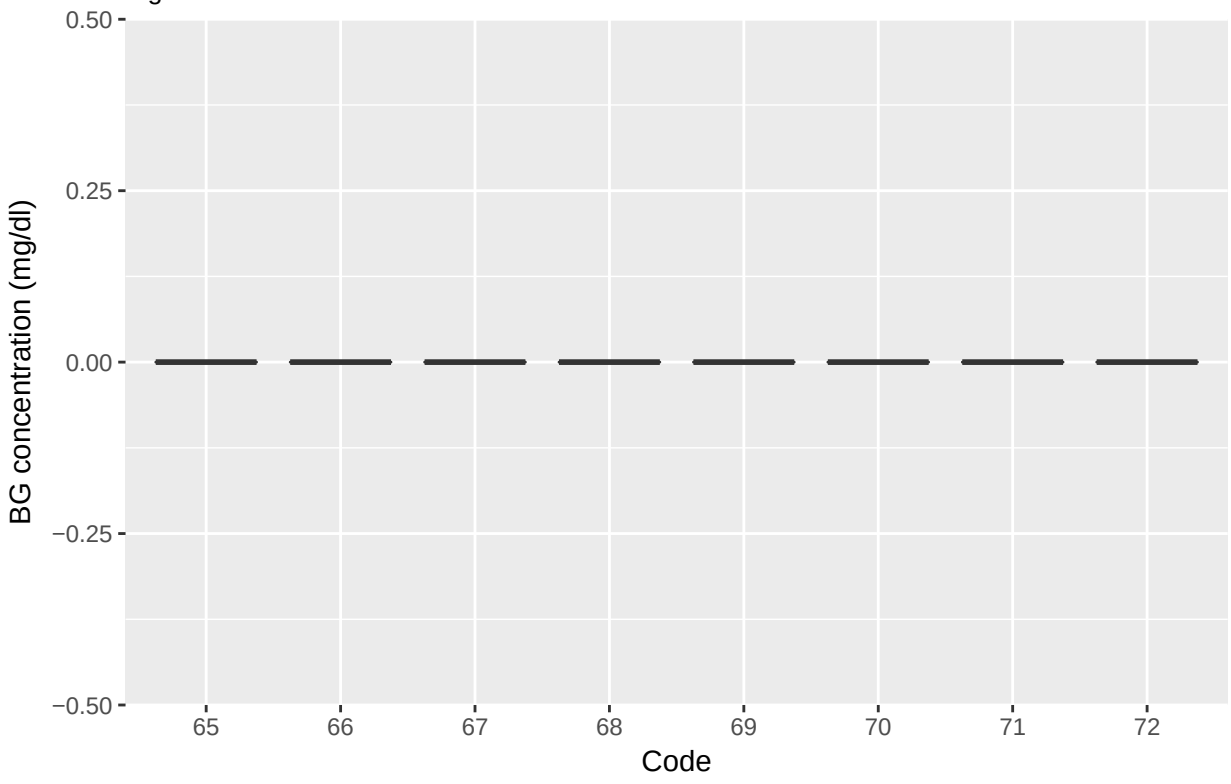
Distribution of measurements based on Pre/ Post Meal

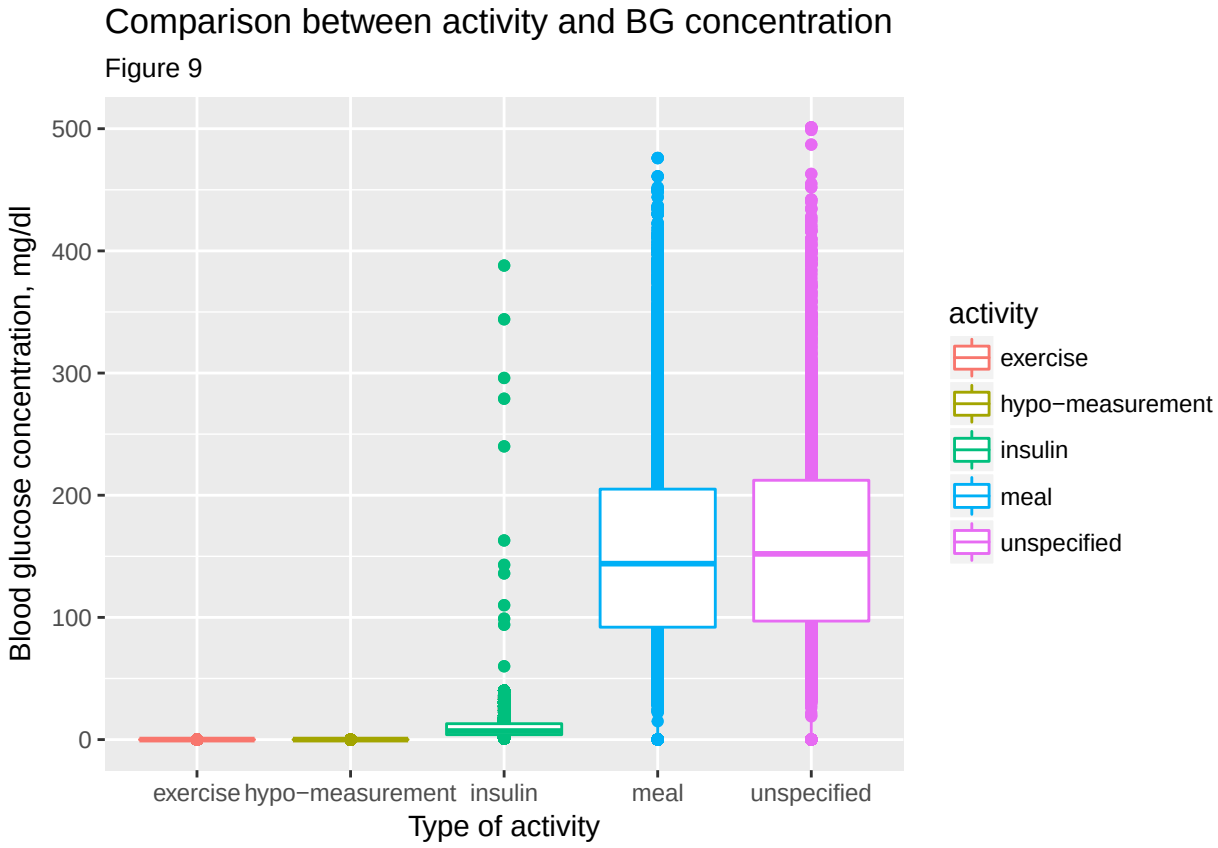
Figure 7



Relation between exercise and Blood Glucose concentration

Figure 8





Conclusion 1: Exploratory Data Analysis

Code vs Blood Glucose concentration

- From Figure 1, we can clearly see a pattern between BG concentration and the type of activity. The patients suffer from Hyperglycemia before or after a meal.
- Also there are three groups which show the similar patterns. To explore more, the three individual plots have been created. Statistical summary of which is as follows
- Figure 4, Relation between Insulin dose and Blood Glucose concentration

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   7.000   9.643  13.000 388.000
```

- Figure 6, Relation between Meal and Blood Glucose concentration

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    97.0   149.0   160.2   210.0   501.0
```

- Figure 8, Relation between exercise and Blood Glucose concentration

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

- Note that, patients take Regular insulin dose more often than NPH and Ultralente insulin dose and rarely measure BG after a meal (Figure 5 & 7)
- Figure 9 nicely compares between the types of activities and BG concentration. Blood glucose level decreases just after taking the isulin dose. It even drops down further down after the exercise. However, the BG notably increases after a meal with a median of around 150 mg/dl

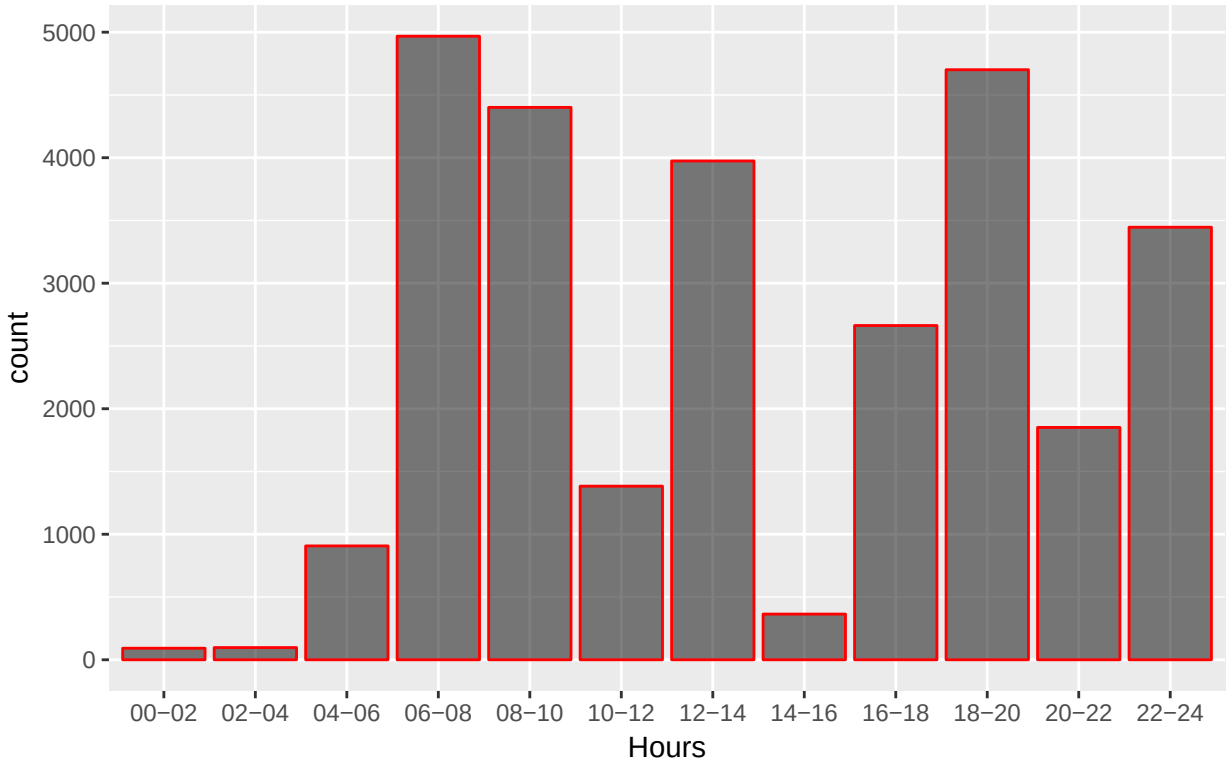
Time vs Blood Glucose concentration

Since time is a continuous variable, it will be difficult to explore the data when time is plotted on x- axis. Which is why the time intervals of 2 hours (time_bin) were created

We want to explore the distribution of Hypoglycemia, Normal BG concentration and Hyperglycemia based on time intervals.

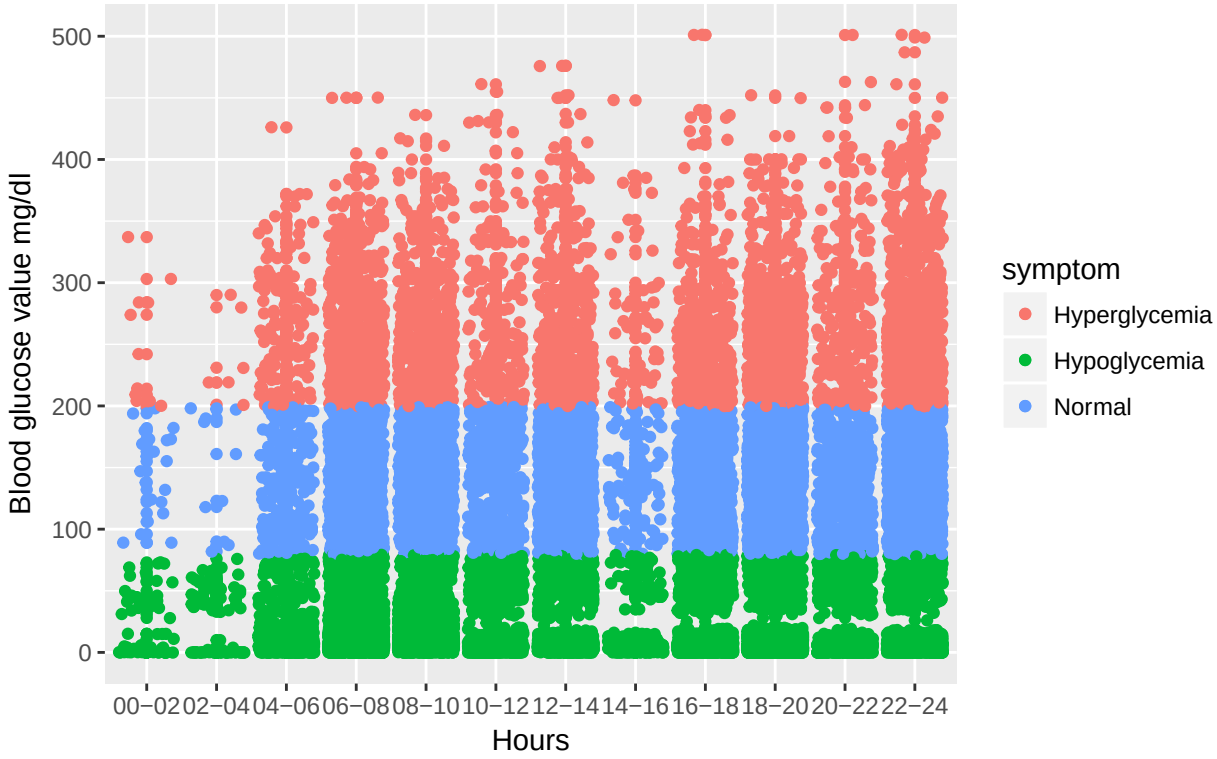
Distribition of BG measurements across time

Figure 10



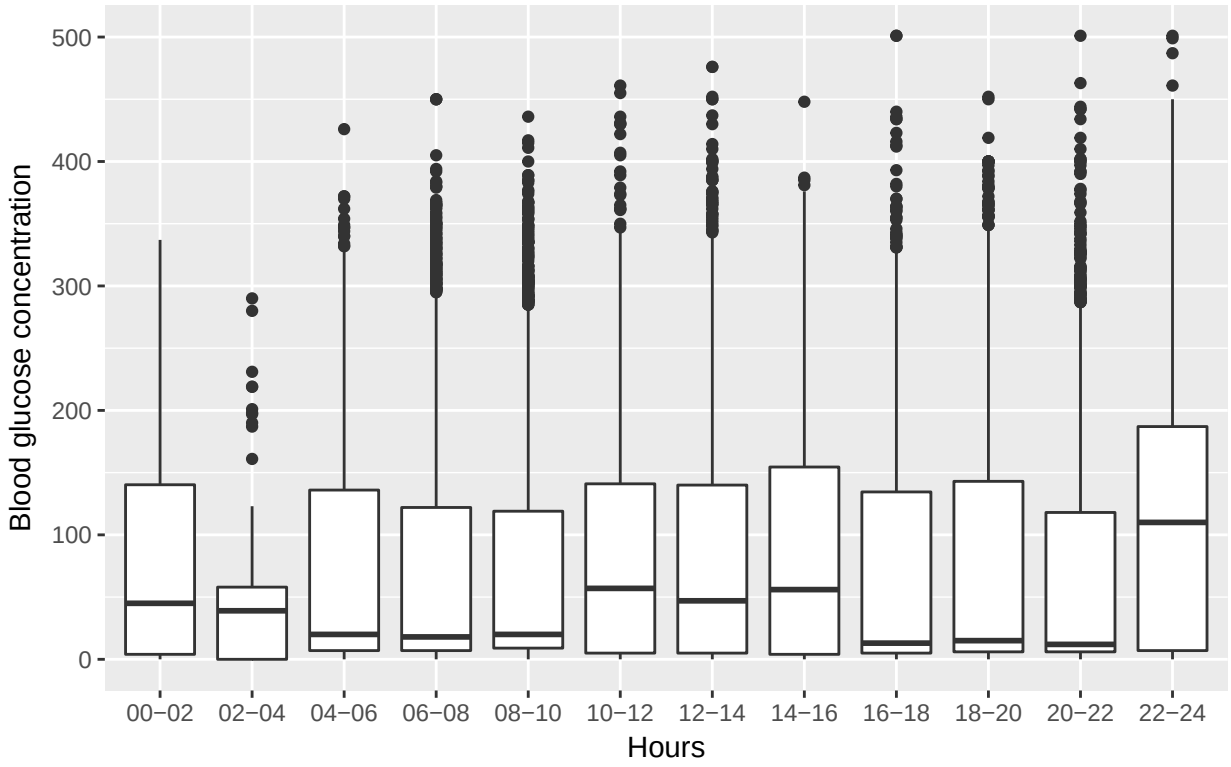
Distribution of BG symptoms over 24 hours

Figure 11



Association between time and BG concentration

Figure 12



Conclusion 2: Exploratory Data Analysis

Time vs Blood Glucose concentration

Each point here represents number of Blood glucose measurements by the patients in 24 hours for several weeks or months.

We do not see any precise time at which the patient was particulary showing Hypoglycemic or Hyperglycemic symptoms, as the symptoms are distributed across 24 hours. However we can say that there are comparatively less measurements showing these symptoms from 00 - 04 in the morning.

This could be due to the fact that there were less number of measurements taken at these time intervals.

Howerver, there is a notable drop in median of BG concentration from 4 - 10 and from 16 - 22.

Regression Analysis

As discussed in the objective of this project, we will be using Logistic Regression classification algorithm to predict Hyperglycemia (high blood glucose) or Hypoglycemia (low blood glucose) in patients.

These factors play an important role in blood glucose management and are directly associated with diet and exercise.

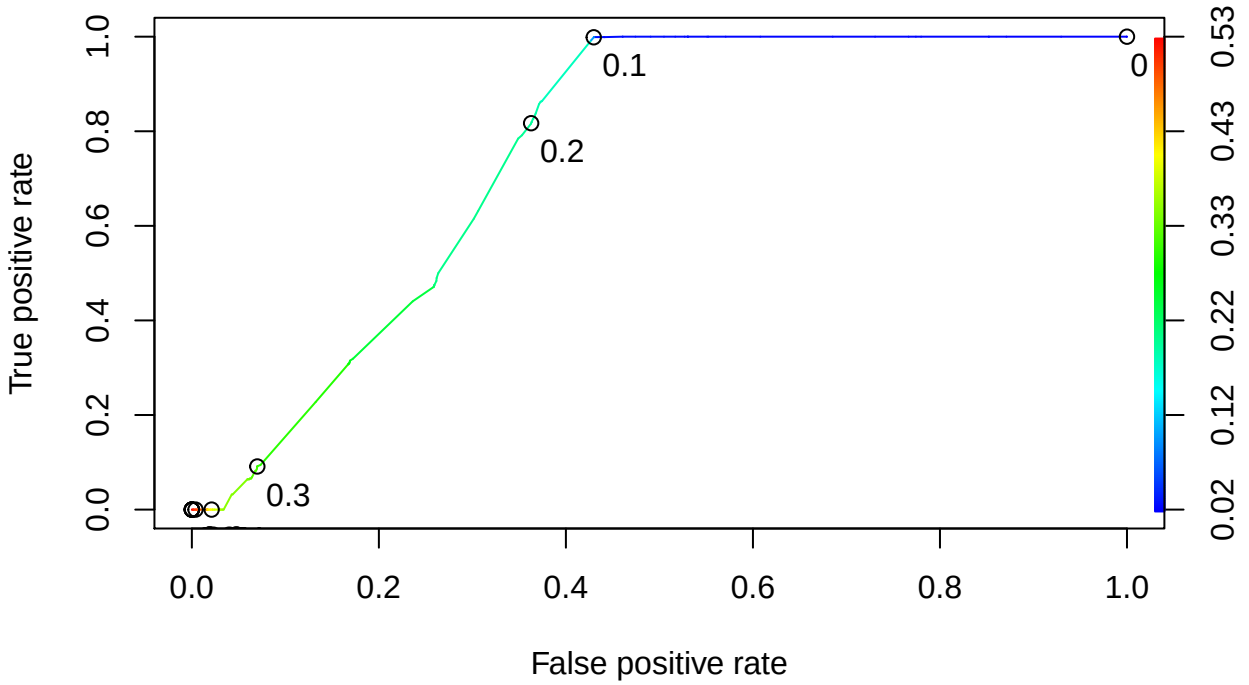
Through predicting the occurrence of these symptoms and the type of activities they are associated with, we can give recommendations to the patient.

Model 1

- In this section we have built a model that predicts hyperglycemic symptoms based on codes
- The codes here represent different types of activities related to meal, exercise, insulin dose etc. (refer to the table from Description of Dataset)
- Before building the prediction model, we have spit the data into train and test dat frames with 70:30 ratio.
- After creating a confusion matrix, the accuracy was taken into consideration to select the model.
- To pick a good threshold value, ROC curve was created

Results

```
##
##     FALSE TRUE
##   0  7474   25
##   1  1156    0
```

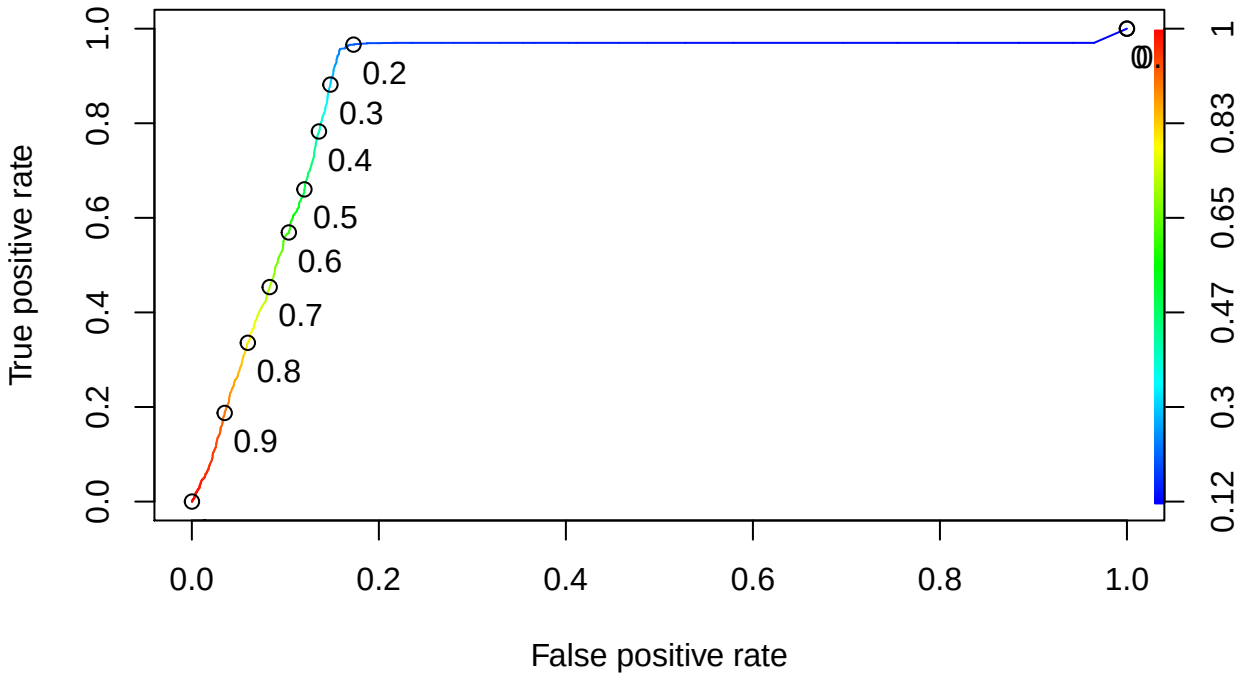


- Our model rarely predicts the risk of hyperglycemia above 50% and the accuracy of the model (86%) was very near to the accuracy of the baseline model (87%)
- Model can differentiate between patients who have probability of suffering from hyperglycemia (AUC = 83%)
- Note that the accuracy of our test is determined by F1 score. The value of which is 0.92 suggesting the strong accuracy of the model.

Model 2

- In this section we have built a model that predicts blood glucose concentration based on activities (Meal)
- Activities are nothing but the similar codes grouped together into 5 bins
  - Insulin (code: 33 - 35)
  - Meal (code: 58 - 64, 66, 67, 68)
  - Exercise (code: 69 - 71)
  - Unspecified (code: 48, 57, 72)
  - Hypoglycemic measurements (code: 65)
- Similar to model 1 we have split the data in 70:30 ratio into a train and test data
- After creating a confusion matrix, the accuracy was taken into consideration to select the model.
- To pick a good threshold value, ROC curve was created

```
##
##      FALSE TRUE
##   0  4679  640
##   1  1136 2200
```



Results

- The model 2 has an accuracy of 79% which is better than our baseline model with accuracy 61%
- The AUC for this model is 88% and the value of F1 score is 0.84 which suggest the good accuracy of the model in prediction blood glucose concentration.

Model 3

- In this section we have built a multinomial model that predicts Hypoglycemia or Hyperglycemia based on code
- We have split the data in 70:30 ratio into a train and test data
- After creating a confusion matrix for multinomial regression, we calculated the classification and misclassification by the model
- Finally a 2-tailed z test was performed to predict the probability of factors associating with blood glucode contration

---

Results

```
##
## p      1    2    3
##   1 1897  912  999
##   2  364 4326  157
##   3    0    0    0
```

- The model 3 has the misclassification rate of only 22%. Which means that the model classifies the test data correctly for 78%

```
##   (Intercept)      code
## 2 0.000000000 0.0000000
## 3 0.004608523 0.3287257
```

- The result of 2-tailed z test shows that the p value of codes 33, 34, 48, & 57 - 64 is almost zero. Hence the confidence level for these codes is very high.

This suggests that the activities represented by these codes are responsible for developing symptoms of Hypoglycemia.

Discussions

- Since the model 1 results were not fantastic, we would not make any reccommendations to our clients based on it.
- However, based on model 2 and model 3 results, we can reccommend the patients to not to take insulin dose right after the exercise as blood glucose is already below normal. And going so will increase the risk of severe hypoglycemia.
- Meal is directly associated with the risk of having hyperglycemia and hypoglycemia. Therefore it is reccommended to eat meals at the similar time everyday to keep track of insulin doses needed per day.

Future Work

- Each type of insulin has an onset of effect, time of peak action and effective duration. This will be explored in detail in future to see which type of insulin is effective in long run.
- In this project we have used only the *internal validation* i.e we have split the data from 70 diabetes patients and predicted the outcome. However, we don't know if the model generalizes to other population. For this purpose a similar dataset will be combined with our dataset to predict the blood glucose level or hyperglycemia and hypoglycemia.
- Similarly, we can combine the data set of healthy patients who are at a high risk of developing diabetes and try to predict the probability with our data set.
