# BLOOD GLUCOSE PREDICTION MODELS FOR PERSONALIZED DIABETES MANAGEMENT

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Warnakulasuriya Chandima Thilina Fernando

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

July 14, 2016

Fargo, North Dakota

# NORTH DAKOTA STATE UNIVERSITY

Graduate School

**Title**

BLOOD GLUCOSE PREDICTION MODELS FOR PERSONALIZED DIABETES MANAGEMENT

**By**

Warnakulasuriya Chandima Thilina Fernando

The supervisory committee certifies that this thesis complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Juan (Jen) Li
Chair

Dr. Simone Ludwig

Dr. Ruilin Tian

Approved:

07/14/2016
Date

Dr. Brian M. Slator
Department Chair

# ABSTRACT

Effective blood glucose (BG) control is essential for patients with diabetes. This calls for an immediate need to closely keep track of patients' BG level all the time. However, sometimes individual patients may not be able to monitor their BG level regularly due to all kinds of real-life interference. To address this issue, in this paper we propose machine-learning based prediction models that can automatically predict patients BG level based on their historical data and known current status. We take two approaches, one for predicting BG level only using individual's data and second is to use a population data. Our experimental results illustrate the effectiveness of the proposed model.

# ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Jen Li for her enormous guidance on conducting a successful research. Her encouragement, effort and advise are the basis of this achievement. I would like to thank Dr. Simone Ludwig and Dr. Ruilin Tian for their valuable opinion on making this research a better one.

# DEDICATION

I would like to dedicate the thesis to whoever would read the thesis and use the ideas of the research to make the world a better place.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# 1. CHAPTER ONE

## 1.1. Introduction

21.0 million is the estimated number of people that have been diagnosed with diabetes, and another 8.1 million people with diabetes have not been diagnosed in the United States in 2012 [3] [1] 1.1. This numbers correspond to 9.3% and 27.8% of total United States population respectively. People who have been diagnosed with diabetes represents 25% of the united stated population over 65years old. It is well known that daily diabetes care is mainly handled by patients and/or their families. The recommendation is to conduct frequent blood glucose (BG) for people with diabetes by their health care professionals in order to achieve a specific level of glycemic control and to reduce the risk of hypoglycemia.

The objective of BG monitoring is to collect detailed information about blood glucose levels at many time points in order to maintain of a more propriate glucose level by more precise regimens. Facilitating the development of an individualized blood glucose profile is being aided by self monitoring of BG levels. This individual profile can then used as a guide by the healthcare professionals in treatment planning for an individualized diabetic regimen. The individualized diabetic regimen is able to give people with diabetes and their families the ability to make appropriate day-to-day treatment choices in diet and physical activity as well as in insulin or other agents. The ability of the patients' recognition of hypoglycemia or severe hyperglycemia can be improved with a personalized diabetic regimen. Further, this will enhance enhance patient education and patient empowerment regarding the effects of lifestyle and pharmaceutical intervention on glycemic control. Individuals can adjust their dietary intake, physical activity, and insulin doses to improve glycemic control on a day-to-day basis by using the adjusted therapeutic regimen.

It has shown that blood glucose self monitoring has yielded improved health outcomes among the patients with type 1 diabetes. A linear correlation exists between the increasing frequency of BG monitoring and reduction of HbA1c among the type 1 diabetes patients [4]. Among patients with type 2 diabetes, an increased frequency of BG monitoring is associated with a better glycemic control among insulin-treated patients who were able to adjust their regimen.

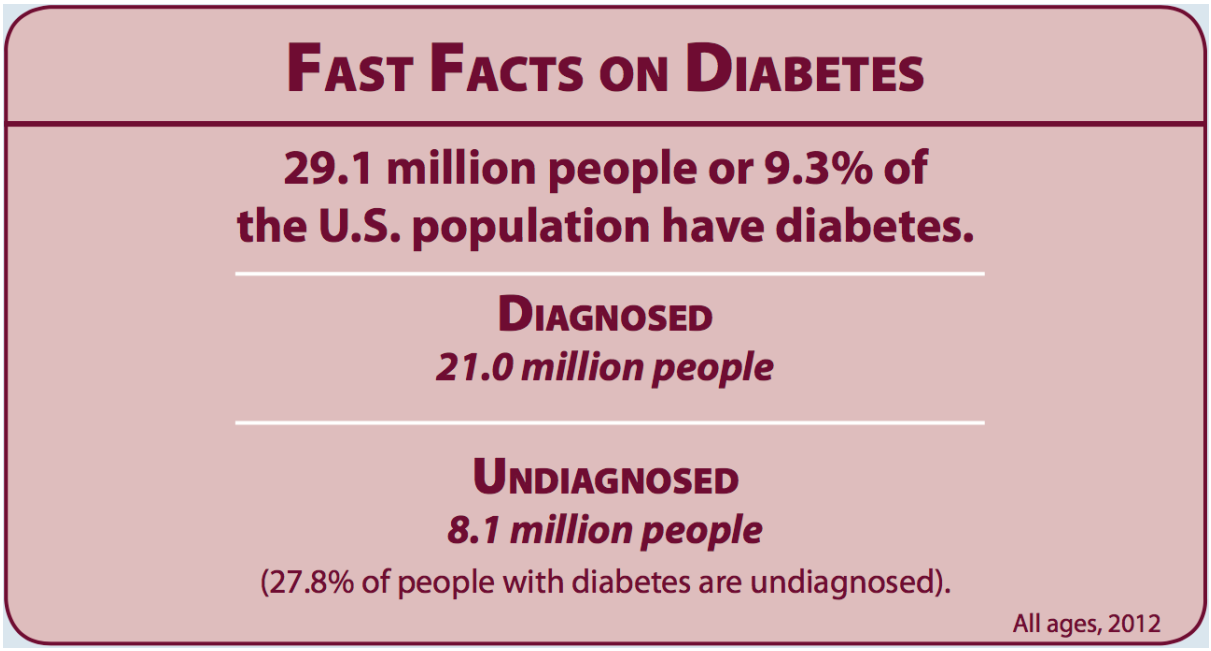


Figure 1.1. Statistic on Diabetes in US [1]

It causes trouble for a person to conduct frequent and regular monitoring of the BG level with the daily activities one has to involve each day, making it impractical sometimes. The frequency with which patients with diabetes should monitor their blood glucose level varies from person to person. The experts recommendation on monitoring blood glucose of insulin treated patients is four times a day. Fasting, before meals, and before bed are the most common times. Under various reasons, patients may not be bale to continue the frequent monitoring of their blood glucose level. For example, patients may be in a conference; may be traveling to a different city or country; may not have the testing devices with them; may not know how to handle the device properly; may forget the test. It's clear that a system to accurately predict blood glucose for self-managing diabetes is essential given that the cumbersome lifestyle of modern people. A prediction system can be used in adjusting the therapeutic regimen answering the change in blood glucose values which in turn help the individuals to adjust their day to day physical activities and dietary intake, and external insulin dose. This will have an improved control over the blood glucose levels of a patient.

Automatic BG prediction using machine learning algorithms has been researched previously as well. For instance, the Artificial Pancreas Project (Juvenile Diabetes Research Foundation 2014), glucose level is predicted so that the insulin flow can be continuously adjusted to meet patient needs

[5]. However accurately predicting blood glucose level for personalized recommendation system still remains a challenge. Most of the existing research has focused on either population based classification or individual based regression. Individual prediction has been carried out based on the data obtained from other patients. It is unknown that if such systems are applicable, accurate, their risk factors and suitability for a personalized recommendation system.

However, individual patient based predictions suffer from an issue called data sparsity problem. In database terminology, data sparsity is know as the number of cells in a data table which are empty. When a patient fails to record a blood glucose measurement it results in an empty value in that patient's record history. Another common issue that individual based predictions would face is the issue of outliers. Outliers can be a result of incorrect data entry, incorrect measurement of blood glucose level, or errors in the measurement equipments.

As for the first part of the experiment, we propose a personalized data-centric predictive model to automatically predict patients blood glucose level. This prediction will be based on patient's daily activities and their historical information. We intend to compare three mostly used machine learning techniques in predicting blood glucose level. Those are support vector machines (SVM), decision trees (DT) and random forest models (RF). We will use three error measurements for calculating the three models' performances. Mean absolute error, root mean square error, and coefficient of determination are the three error measurements. The machine learning technique with the highest accuracy will be considered for the comparison in the second and third parts of the experiment. The second part of the experiment is to use whole population data in predicting blood glucose level. For that, pooled panel regression model will be used. By using panel regression, we intend to overcome the issue of data sparsity. The panel data regression model will be run on the same set of data as used in the first experiment and the results will be compared with least error model from the first experiment. The third section of the experiment will be an enhancement of the pooled panel data regression model where pre clustering of the data is used to identify a similar set of patients and then apply the panel data regression model on the clustered patients. We call this pre-clustering based personalized regression model. We will apply hierarchical clustering to the dataset to extract the patients with similar blood glucose levels. The hierarchical structure will be sliced in different distances to get clusters with variable member size and the panel data

model will be applied to each cluster. The results are then analyzed to determine the effectiveness of clustering.

The rest of the thesis is organized as follows. Chapter 2 presents the related work which has been done in the filed of blood glucose prediction. Chapter 3 describes the research methodology used. Chapter 4 presents the results of the experiment and evaluation of the results. Conclusions are provided in chapter 5 , respectively.

# 2. CHAPTER TWO

## 2.1. Related Work

A study on the literature on diabetes, the use of mathematical models is prominent. On of the pioneer work has been proposed by [6] to capture the glucose-insulin interaction. In his experimental approach, he was able to find a set of coefficients of the relationship between insulin level and the blood glucose level in blood.

The famous minimal model was later proposed by [7] that considers the meal effects. It interprets the complex dynamic plasma insulin response to glucose injection. Variant models have been developed to capture the effects of physical exercise on accelerating the utilization of glucose and insulin, as well as increasing the muscular and liver sensibility to insulin [8]. Their model has been able to compare the blood glucose in normal, non-insulin-dependent diabetes and insulin-dependent diabetes people. In recent year, the the model has been capturing the attention of researchers and practitioners.

There exists various studies in risk prediction related to diabetes management. Early in 1996, Shanker proposed a population based classification model using artificial neural networks to predict the onset of diabetes mellitus among the Pima Indian female population near Phoenix Arizona [9]. They have modeled a relationship between onset of diabetes mellitus and various risk factors for Prima Indian females using the artificial neural networks. The risk factors include number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, body mass index, age and diabetes pedigree function derived by a previous researcher. The performance of their approach is compared with logistic regression model and another model called ADAP. Their results claim that artificial neural network method, with 81.4% accuracy on the test data, is a better option for classifying onset of diabetes over logistic regression and ADAP model. However our research focuses on predicting the exact blood glucose values rather than performing a classification of patients to determine their risk of being a diabetic or not.

More recently, in the Artificial Pancreas Project (Juvenile Diabetes Research Foundation 2014), glucose level is predicted so that the insulin flow can be continuously adjusted to meet patient needs [5]. The author points out one major issue in the software in the artificial Pancreas

which determines the amount of bolus and basal insulin needed to respond to the changes in glucose levels between meals and rapid rise in glucose levels during the mealtime. The issue is that there is no sufficiently robust model for predicting the necessary doses of insulin in various situations. however work is aimed on predicting the blood glucose level given the factors which would affect the blood glucose level.

Zecchin et al. quantified the potential benefits of glucose prediction in terms of reduction of number/duration of hypoglycemia [10]. In his work, he has used artificial neural network to predict the glucose concentration of the blood after meal intake. A pool of 10 patients has been the subject for the training set. He has been able to achieve a root mean square error of 16.6. However he has only used the blood glucose values from continuous blood glucose monitoring. The meal information, external insulin information, physical activities, stress, emotions and other factors which would affect the the blood glucose level have not been taken into consideration.

In predicting blood glucose level, numerous machine learning algorithms have been applied. One such study has been done by Marling et al and their work is focused on measuring the glycemic variability [11]. Measuring glycemic variability has been studied by the early researches and using standard deviation, mean of daily differences, continuous overlapping, net glycemic action over a certain time period and daily risk range. The authors have used a machine learning algorithm to incorporate physician perception into existing continuous glucose monitoring systems. They have used an individual based approach and two physicians have determined if each patient exhibits excessive variability in their continuous glucose measurement chart. That classification knowledge was incorporated into a machine learning classifier implemented using Weka.

Artificial neural network (ANN) and neuro-fuzzy systems were proposed to use to predict blood glucose level for expert management of diabetes mellitus by Patterson and Sandham [12]. Authors have selected patient's diet information, exercise, insulin regime data and another variable for factors such as stress, illness, pregnancy, etc. for training the artificial neural network. Their previous blood glucose levels along with the above parameters have been used to produce an optimum therapy using a neuro-fuzzy optimizer. This training has been done for each patient and artificial neural network weight matrices have been obtained for each patient. Those are called predictor weight matrix and optimizer weight matrix and they are used for predicting the blood glucose level of that particular patient. The back-propagation algorithm incorporating momentum

and adaptive learning rates were used to train the algorithm. Two activation functions were used; one for the neurons in the recurrent layer and another for the neurons in the output layer. However the paper does not contain a numeric measurement on how well their prediction is compared to the actual blood glucose values of the patient. And they have been able to test the method only on two patients. Our approach uses data on a group of similar patients in predicting the blood glucose level where the approach of the paper is individual based. On the process of gathering data for the experiment from the two patients, one of the patient was not able to record data from a patient for two days. Authors are concerned about the issue of missing data and they are hopeful of a solution to the issue in future by some other researchers which motivated us to use population data to overcome the missing data problem.

In a study by Karim Al Jabali, artificial neural networks were used to create dynamic simulation model of type 1 diabetes, that simulates the progression of the disease and the two term neural controller that is responsible for the insulin released to stabilize the glucose level [13]. The neural controller mimics the pancreas section of insulin into the body. Their dataset used for training the artificial neural network includes previous blood glucose level, short term, mid term and long term insulin data, exercise data and meal information. They have been able to achieve a deviation of 3% to 12.4% of the actual blood glucose level to the predicted blood glucose level. The results shows that the use of complex neural network architectures could effectively emulate the working of controllers that deliver insulin to Type 1 diabetic patients.

Plis et al. proposed a generic physiological model of blood glucose dynamics to generate informative features for a Support Vector Regression model to predict blood glucose levels [14]. They have eight features as their dependent variable; carbohydrate consumption, carbohydrate digestion, subcutaneous insulin, insulin mass, level of active plasma insulin, blood glucose mass and blood glucose concentration. The approach the researches have followed was individualized where a regression line was modeled for each patient. The support vector regression model was compared with two other models. The first model, the baseline model, assumed the initial blood glucose level at time zero does not change over time. The second model is an auto regressive integrated moving average model (ARIMA). Based on the results of the experiments, support vector regression model outperforms both the baseline model and ARIMA model where support vector regression model records a root mean square error of 35.8 and ARIMA and base line models

record errors of 39.6 and 41.7 respectively. Particular results achieved by the research helped us to focus on other regression models to predict blood glucose level and compare such models against the support vector regression model.

Decision trees have been used to identify patients with diabetes and to predict a diabetes patient. Pociot et al. proposed a novel analytical method to predict type 1 diabetes mellitus [15]. Their research includes both decision trees and artificial neural networks to analyze type 1 diabetes mellitus genome-scan data. Their objective was to identify markers in a particular genome which could characterize a diabetes or non diabetes status. The study has shown that both the artificial neural network approach and decision tree approach are good in identifying markers, hence they can be used to predict patients with diabetes. Han et al. used decision trees to build a model for diabetes prediction from Pima Indians Diabetes Dataset [16]. The Prima Indian Diabetes Dataset consists of 8 types of data related to pregnant women. They include the number of times of pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness. 2-hour serum insulin, body mass index, diabetes pedigree function and the age of the patients. The authors have used a tool called RapidMiner to construct the decision tree model. The decision tree model has helped to derive that Plasma-Glucose is the main attribute that leads one to know whether a patient will develop diabetes. They have been able to achieve 72% of accuracy by using decision trees.

The two studies mentioned above provided a sound platform for using decision trees in predicting blood glucose level as the algorithm has shown promising results in classifying a patient in to a diabetes or non diabetes. A study has been conducted by Sudharsan et al. in order to determine the accuracy of various machine learning techniques in predicting blood glucose levels [17]. The dataset of the experiment consists of 56000 self monitored blood glucose values from pre-identified patient data from clinical trials of patients with type 2 diabetes. The dataset of the research was limited only to blood glucose values. Other factors which affect patient blood glucose levels have not been taken into consideration. The predictions are made for each individual, hence this study can be categorized as an individual approach in predicting blood glucose levels. The machine learning algorithms of consideration were k-nearest neighbor, support vector machines and naive Bayes classifier. The results of the experiment shows that the highest accuracies have been

achieved by support vector machine algorithm of which the accuracy exceed 90% where k-nearest neighbor and naive Bayes algorithms have recored an accuracies below 50%.

Although many researches exist, the task of predicting glucose level and make personalized recommendations still remain to be a challenging task. The population based approach was the basis for most of the existing researches. In population models, the prediction of one patient's situation is based on all of data of the population. Whether they are applicable or accurate for individual risk estimations or personalized recommendations is unknown. Our our focus moves from population based analysis and prediction to individual based analysis and prediction.

# 3. CHAPTER THREE

## 3.1. Methodology

The first part of the section describes the dataset used and the preprocessing carried on the dataset for the experiment. Next the three main parts of the experiment; time series prediction on individual data, pooled panel regression model and the pre-clustered panel data regression model; are explained.

### 3.1.1. The Dataset

The experiments were carried out using the diabetes dataset for 70 patients from the UCI Machine Learning Repository **??**. The dataset includes the information about the blood glucose level, insulin intake, diet and exercise for the patients at different times for varying time periods for the patients. There are 70 files, each containing one patients records. Diabetes files consist of four fields per record: time, date, code and value corresponding to code.

File Names and format: (1) Date in MM-DD-YYYY format (2) Time in XX:YY format (3) Code (4) Value

The code field contains the codes for a patients blood glucose level, insulin dose, level of exercise and diet etc. The corresponding entry in the value field gives the value for the specified code in the record. The Code field are matched to their meaning and can be found in table **??**.

The first preprocessing task was to arrange the data in chronological order. However, the timestamp recorded in the dataset when a blood glucose measurement is taken or when a meal or exercise or external insulin information is entered, is not consistent among all the patients. For a meaningful comparison of the prediction results, we converted the local timestamps in each of the record to a common timestamp.

In order to identify a common a timestamp for all records, we extracted the hour from original timestamps of all the records in 70 patients. Then the frequency of each hour was plotted against the hour to build a histogram. The histogram is shown in figure.

analyzing the histogram, we were able to identify the time periods when most of the the records have been made. This information was used to define 4 categories for the timestamps. The

Table 3.1. Dataset code description

| Code | Description |
|---|---|
| 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 |

categorization can be found in table. The respective category number number was assigned based on the time of the record.

In the dataset, both the dependent variable and the independent variable were recorded together along with the code. Codes 48, 57, 58,59, 60, 61, 62, 63, and 64 from the table are the features or the dependent variable and they were assigned to a different columns with respective

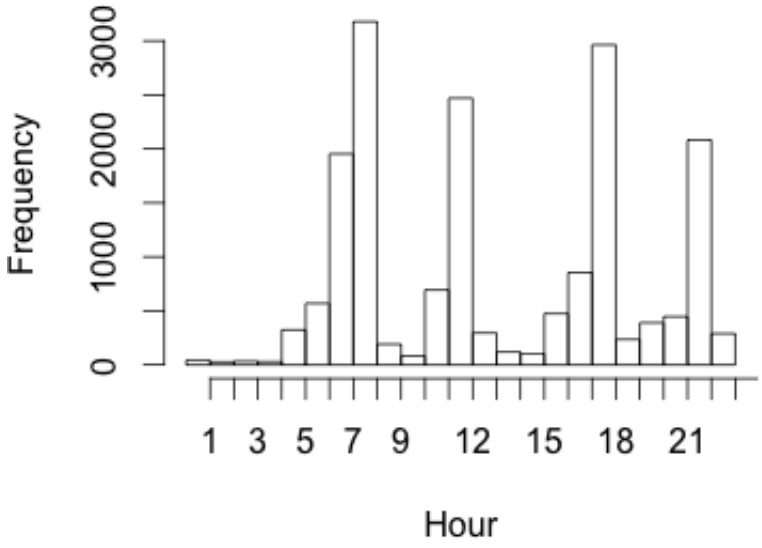


Figure 3.1. Histogram for number of records in each hour

Table 3.2. Timestamp categorization

| Timestamp Hour | Time Category |
|---|---|
| 00:00 - 09:00 | 1 |
| 09:00 - 14:00 | 2 |
| 14:00 - 19:00 | 3 |
| 19:00 - 00:00 | 4 |

to the time they were recorded. The remaining codes, 34, 35, 65, 66, 67, 68, 69, 70, 71 and 72, contained the blood glucose measurements recoded at different times of the day. A new column was created for the independent variable, which was the blood glucose measurement and they were arranged with respect to the time they were recorded.

The original dataset records contained alphanumeric values instead of the numeric values. Those may have been a result of incorrect data entry. Such instances made the records invalid and they were eliminated from the dataset.

After completing the pre-processing, the dataset contained 12675 records on unique timestamps combining all the patients.

### 3.1.2. Time Series Predicting Model

The idea of the time series predicting model is to analyze the performance of three machine learning algorithms in predicting blood glucose levels. The patients' historical blood glucose measurements, which are taken at different different time intervals are needed to predict the patient's

blood glucose measurement. The time intervals corresponds to the daily activities which have the highest impact for the blood glucose levels. The events are external insulin intake, meal intake, exercise and sleep.

The prediction problem we used is the time series forecasting. The main reason is that the data in of BG level follow a timely pattern. A time series is a sequence of data points, typically consisting of successive measurements made over a time interval. Time's natural ordering to the observations marks the main different of a time series dataset to a regular dataset. Another unique feature of these datasets is that adjacent observations are dependent (Box et al., 2008).

We limit our study only to time series prediction. In the time series regression problem, basic platform is to use the BG measurement at time t and the other dependent variables to predict the BG level in time t+1. The following equation shows the basic formula.

$$BG_{t+1} = f(BG_t, BG_{t-1}, BG_{t-2}, ..., BG_{t-n}, X_{1t+1}, ..., X_{nt+1}) \tag{3.1}$$

In the dataset with n points, t refers to the current or the most recent observation. The last observation is t - n. $X_1$ ... $X_{nt}$ refers to the features which causes the change in blood glucose levels at time 1 to nt. Function f is known as the model which estimates the blood glucose value for a future time stamp.

In our work, we have implemented time series predicting with three different models, namely support vector regression model, decision tree regression model and random forest model.

**3.1.2.1. Support Vector Regression**

Support vector regression model is the first implementation that we used to predict the blood glucose measurement [2]. The main idea of Support Vector Regression (SVR) is that, given a set of training data containing a feature vector, $X_1$, $X_2$, ... , $X_n$ and the output Y, a function f(x) has to be derived such that the function has less than $\varepsilon$ deviation from all the Y outcomes(cite the image here). In the diabetes dataset, $X_1$, $X_2$, ... , $X_n$ are the patient's insulin intake, exercise, meal intake and etc. The Y outcome is the blood glucose measurement at a particular time. The following is the function of f(x).

$$f(x) = \langle w, x \rangle + b \quad with \quad b \in \mathbb{R} \tag{3.2}$$

In support vector regression, it assigns a normal vector w and denotes the dot product of x in w. A small w value will ensure the function f(x) is flat along the x axis. Hence the problem can be transformed to an optimization problem. In order to keep the optimization problem feasible, some error value is allowed in the model.

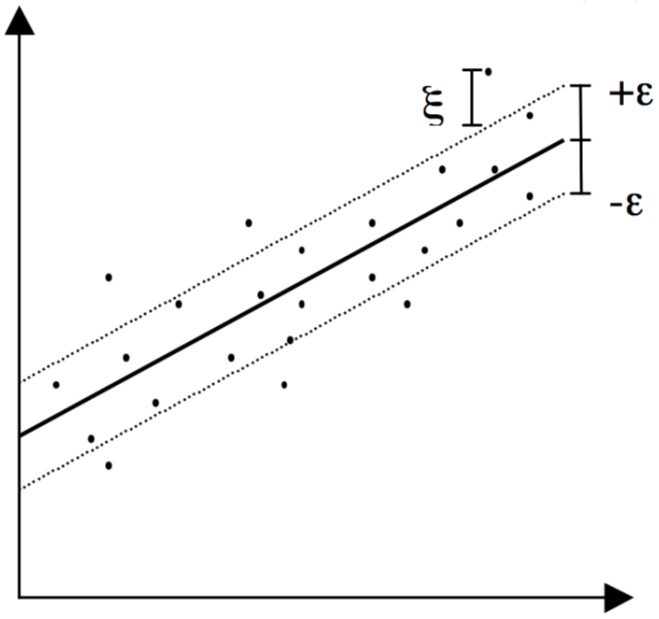


Figure 3.2. Pre-specified accuracy epsilon and a slack variable $\xi$ in SV regression [2]

#### 3.1.2.2. Decision Tree Regression

Decision Tree Regression is another method for time series prediction. We used it to predict the blood glucose level. In such regression models, a target variable does not contain a class. The basic idea is to fit a regression model to the target variable, which is the blood glucose level in our experiment, by using the independent variables [18]. Patient's insulin intake, exercise, meal intake and etc. will be the independent variable in our experiment. In decision tree regression model, for each independent variable, data is split at various split points. Finding a good partitioning of the data for splitting can be achieved by maximizing the information gain at the target variable Y which is the corresponding blood glucose value.

The decision tree model starts with a single tree which contains all the variable. Then it calculates both the sum of square error and the prediction for each leaf. It will search over all the binary splits of all the independent variables to find the one which gives the least sum of square

error. If the error decrease is less that a threshold, it would stop the model. Otherwise it would create another two new child nodes and repeat the process.

#### 3.1.2.3. Random Forest Regression

The random forest model follows the usage of many decision trees for regression. Breiman at el have mentioned in his paper that a significant advancement can be achieved in classification accuracy when a number of decision trees are generated and by letting the trees vote for the most popular class [19]. The idea is to generate a set of random vector which governs the formation of each decision tree.

One way of forming the tree is to make random selections without replacement from the training set and form the tree. Another way is to make random split selections where at each node, the random split os selected at random from a K best splits. Another way is to use a random set of weights to make the selection from the training set.

However the idea of the random forest regression by Breiman at el is to generate a random vector for $\Theta$ for each k trees where k has to be defined prior to initiating the algorithm. The random vector $\Theta$ k should be independent of all the past $\Theta$ k-1 vectors, however all should follow the same distribution. A tree is formed by using the $\Theta$ k random vector and the training set to form one classifier for regression line, h(x,$\Theta$ k). x would be the set of independent variables.

Each split in each tree will be decided based on the random number in the random vector. After generating all the trees, each tree gets the chance to make a prediction and the final predicted value of the random forest regression model is the average prediction of the predictions made by all trees.

It is important to identify the number of trees k in the random forest model which would yield the least error. Hence, we will vary the tree size of the forest and generate the error rate and plot the error against the number of trees. That optimal number of trees will be used when random forest model is used to compare with any other model.

### 3.1.3. Pooled Panel-Data Regression Model

The orthodox time series models use an individual's historical data to make the prediction. It calls for a data collection where a large amount of data to be gathered on an individual for whom we are going to do the prediction. However there can be instances where the individual's own data may not be sufficient to make a prediction. One situation may be that a new patient is observed

where he does not have any historical data. This issue can be addressed by introducing the pooled panel data regression model for diabetes prediction. The main advantage of this approach is that it eliminated the problem of data sparsity by using the cross patient information as the model estimates predictions by using the historical information of all patients.

Prediction on panel data is made by pooled regression. As suggested by existing studies, the inability to predict with time-series regression models is a small sample issue [20]. Therefore, we used panel data to increase sample size. As proven by the researches of [21], [20], and [22], prediction models built on pooled regression models estimated on panel-data dominate those of time-series regression forecasts.

The linear regression model was used to predict the pooled panel model to predict the blood glucose levels. The linear regression models the function that maps the dependent variable which is the BG level and the independent variable denoted by X. This mapping is modeled using linear predictor functions whose unknown model parameters are estimated from the data. Let BG denotes the dependent variable whose values we wish to predict, and let X1, ... , Xk denote the independent variables from which we wish to predict it, with the value of variable Xi in period t (or in row t of the data set) denoted by Xit. For example, in our dataset, Xi can be insulin dose, hypoglycemic symptoms, meal ingestions, and exercise activity. Then the equation for computing the predicted value of BGt is:

$$BG_{\mathrm{t}} = b_0 + b_1 X_{1\mathrm{t}} + b_2 X_{2\mathrm{t}} + ... + b_{\mathrm{k}} X_{\mathrm{kt}} \tag{3.3}$$

The prediction for BG is a straight-line function of each of the X variables will be carried out by holding the other variables fixed, and the contributions of different X variables to the predictions are additive. The slopes of their individual straight-line relationships with BG are the constants b1, b2, ... , bk, the coefficients of the variables. That is, when keeping the other variables a constant, bi is the change in the predicted value of BG per unit of change in Xi. The additional constant b0, the so-called intercept, is the prediction that the model would make if all the X's were zero (if that is possible). The coefficients and intercept are estimated by least squares, i.e., setting them equal to the unique values that minimize the sum of squared errors within the sample of data to which

the model is fitted. And the model's prediction errors are typically assumed to be independently and identically normally distributed.

#### 3.1.4. Pre-clustered Personalized Regression Model

In the third part of the experiment, we aimed to verify that pre-clustered regression would improve the prediction accuracy by providing more personalized prediction. Of course, like pooled panel-based prediction, this approach would also help to remedy the data sparsity problem faced by individual data.

The underlying model's heterogeneity affects the performance of a pooled panel data model. For improving the performance and minimizing the affect of the issue, we segment patients into groups whose prediction models are quite similar. we cluster patient data to group similar patients. The similarity of patients is defined by the similarity of the blood glucose functions. The lesser the difference between the blood glucose values the similar the patients are. The goal of clustering is to remove the influence of patients who have little or no similarity with the testing patient for whom predictions are being made. Clustering is based on similarity between patients, so that patients who are quite different from the testing patient are removed from his/her cluster. One main advantage of clustering is that patients in different clusters, which are dissimilar from another cluster, do not affect the prediction score of the testing patient. Thus, removing these patients does not result in loss of information, but effectively reduces noise caused by these data. In order to get the similarity between the patients, an appropriate similarity measurement should be applied. The measurement should be able to distinguish the patients based on a chosen criteria.

We apply hierarchical clustering to group patients based on their similarity. Our data sets did not include sufficient patient features except for those specified by the code field, such as blood glucose level, insulin dose, level of exercise, and diet. Due to the limitation of the data set, we can only measure the similarity based on these features patterns. We choose the blood glucose level as the candidate feature pattern making the assumption that the patients with close blood glucose values at each category have a similar pattern in blood glucose variation.

We used the Euclidean distance between patients' blood glucose level at each time category. We observed in the dataset that the number of blood glucose measurements differ from patient to patient. In order to have a uniform distance measurement and a comparison across all the patients, we decided to choose the first 200 measurements. 4 patients were discarded as they did not have

more than 200 records for blood glucose values. The new dataset contained records of 67 patients. The Euclidean distance between patient p and q with respect to each of the n time categories are shown below.

$$distance\,(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_{\mathrm{n}} - q_{\mathrm{n}})^2} \tag{3.4}$$

A pairwise distance matrix is made based on the pairwise distances between all patients' blood glucose values. The pairwise distance matrix is used to construct a dendrogram, a tree diagram representing the similarity among the entities. The dendrogram is then used to perform the hierarchical clustering. We have used the agglomerative method to build the hierarchical clusters. The hierarchical cluster which was obtained for the 70 patients is shown in Fig. 3. We can slice the dendrogram across the distance axis to separate patient clusters

We had to stick with the hierarchical clustering based on blood glucose level because of the limitations of the Diabetes Dataset provided by UCI Machine Learning Repository. A dataset with more patient information, for instance patients' age, sex, marital status, occupation ethnic group, type of food consumed could have been used to perform k-means clustering or other relevant clustering technique. Such clustering would be useful in identifying patients with similar pattern with regarding to age or sex which will be more meaningful.

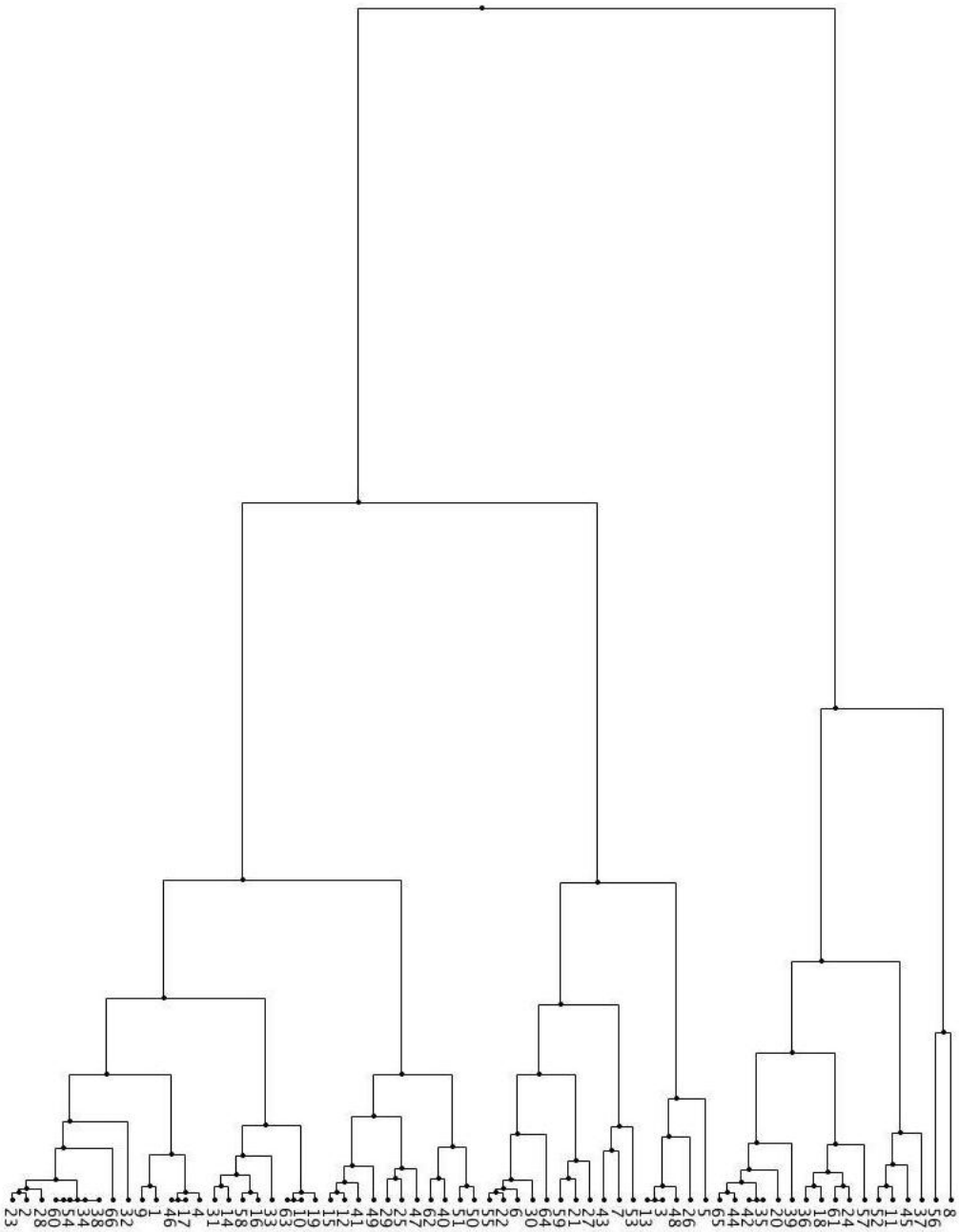


Figure 3.3. Dendogram for 67 patients with respect to their similarity distance

# 4. CHAPTER FOUR

## 4.1. Evaluation

This section describes the experiment results and the evaluation of the results. First the error measurements which are used to evaluate the results are described. Next the three models, support vector machines, decision trees and random forest, used for individual regression analysis are compared. The third subsection describes the results obtained in pooled panel regression model. Final subsection will illustrate the findings from pre-clustered panel data regression model.

### 4.1.1. Error Measurements

We use three evaluation criteria namely Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and coefficient of determination.

#### 4.1.1.1. Mean Absolute Error

The mean absolute error is the average of error calculated by the predicted blood glucose level and the actual blood glucose level [23]. p is the predicted blood glucose level and y is the actual blood glucose level. n is the number of instances where a blood glucose measurement is taken.

$$MAE = \frac{1}{n}\sum_{i=0}^{n} \mid p_{\mathrm{i}} - y_{\mathrm{i}} \mid \tag{4.1}$$

#### 4.1.1.2. Root Mean Square Error

The root mean square error measures the standard deviation of the differences in the predicted blood glucose level ($p_{\mathrm{i}}$) and the actual blood glucose level (y) [24].

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n} (y_{\mathrm{i}} - p_{\mathrm{i}})^2}{n}} \tag{4.2}$$

#### 4.1.1.3. Coefficient of Determination

The coefficient of determination is a heavily used in statistic. The coefficient of determination, denoted as $R^2$, measures the proportion of the variance of the blood glucose level, which is the dependent variable, that can be predicted from the given features which are the independent

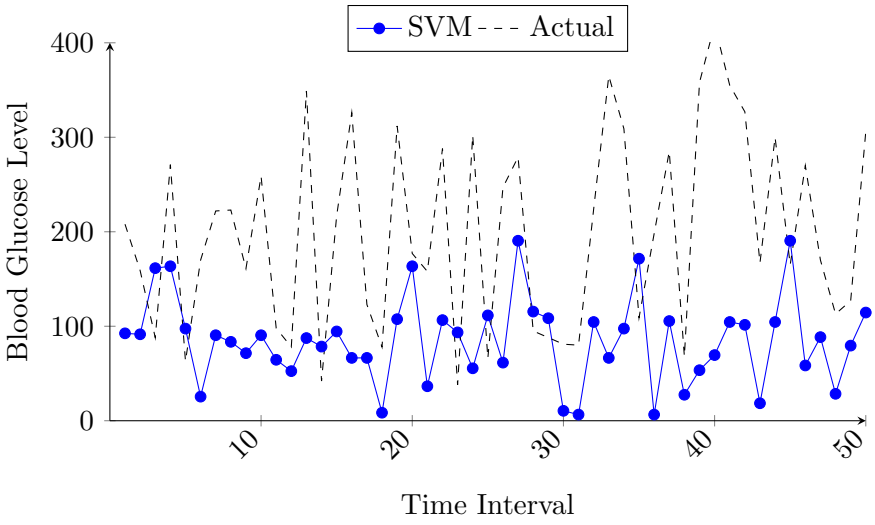


Figure 4.1. Prediction of blood glucose level by support vector regression

variables [25]. This is an indicator to show how well the data is fit to the model. The explained sum of squares is the sum of the squares of the differences of the predicted values and the mean value of the response variable, y, the blood glucose level.

$$R^2 = 1 - \frac{Sum of squares of the residual}{explained sum of squares} \tag{4.3}$$

### 4.1.2. Individual Prediction Model

The first set of experiments were carried out to analyze the performance of individual-based regression using the 3 time series regression models.

The first model we used was the SVM model. The figure is drawn for the first 50 blood glucose level predictions and the actual blood glucose levels for a randomly selected patient 4.2. . y-axis is for the blood glucose level and the x-axis is for the timestamp. ( Day 1 category 1 is x=1, day 1 category 2 is x=2, day 2 category 1 is x=3, day 2 category 2 is x=4 ... ). The second model we chose to analyze is the decision tree model. Similar to the SVR model, we plotted the predicted blood glucose values against the actual blood glucose values 4.2.

The third model we chose to analyze is the random forest model. We conducted an experiment by calculating the root mean square error yielded by the random forest model when the number of trees are increased from 1 to 250. The error graph 4.3 shows that the least error is achieved when the number of trees are 68 and the error rate does not decrease afterwards. Similar

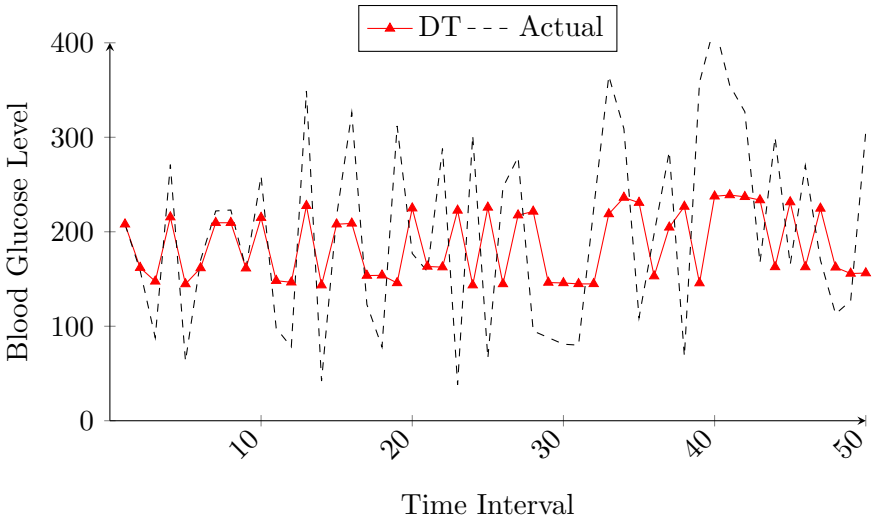


Figure 4.2. Prediction of blood glucose level by decision tree regression

to the two previous, we plotted the predicted blood glucose values against the actual blood glucose values 4.4. The three models were plotted in the same graph in figure 4.5.

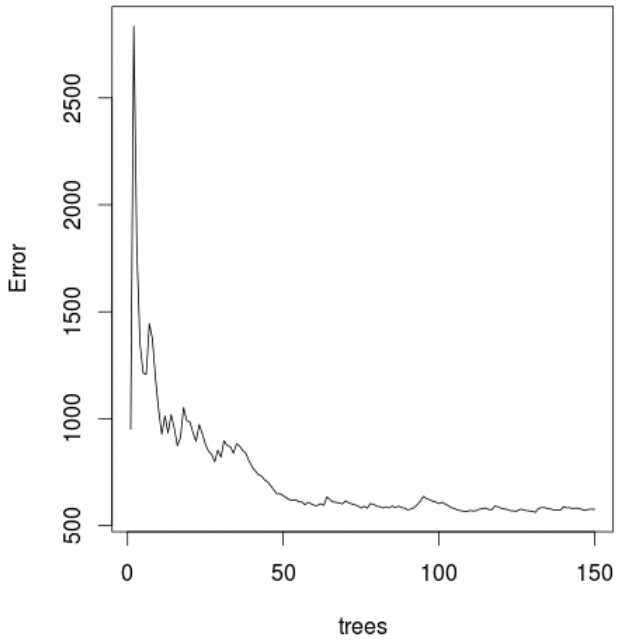


Figure 4.3. Root mean square error change against the number of trees in random forest regression

.

As illustrated in the figure, the accuracy of the SVM-based approach was not ideal. Although it tends to follow the pattern of the actual blood glucose level, the variance between the

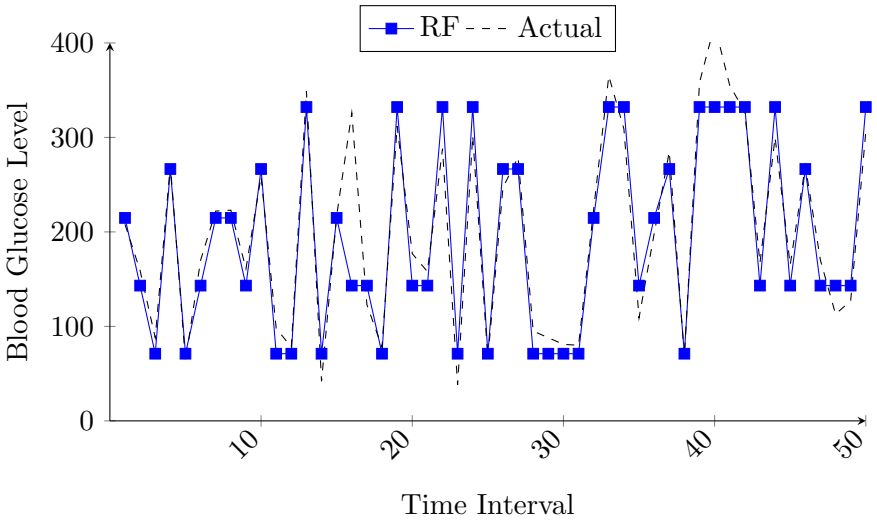


Figure 4.4. Prediction of blood glucose level by random forest regression

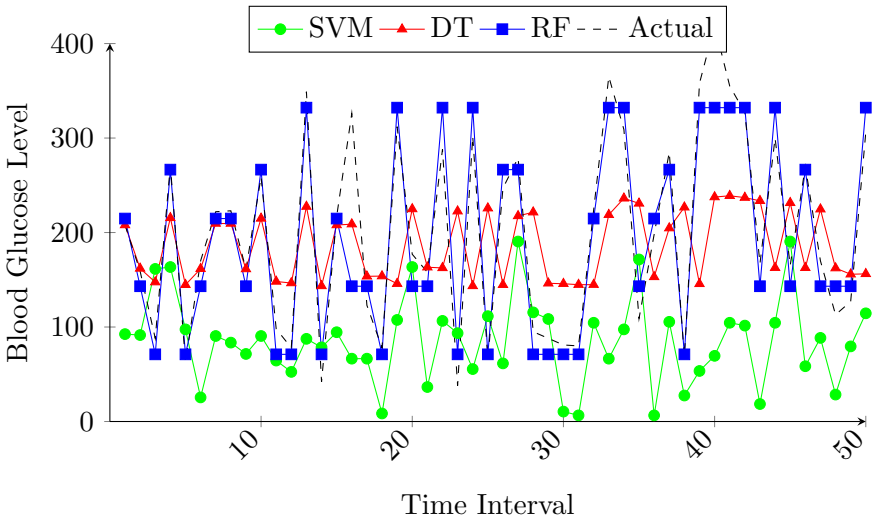


Figure 4.5. Prediction of blood glucose levels using SVM, DT and RF methods

Table 4.1. Averages error metric over all patients for each individual time series regression model

| Error Matric | SVM | Decision Tree | Random Forest |
|---|---|---|---|
| RMSE | 68.76 | 41.06 | 39.73 |
| MAE | 63.097 | 36.423 | 37.586 |

Table 4.2. $R^2$ over all patients for each of the individual time series regression model

| Error Matric | SVM | Decision Tree | Random Forest |
|---|---|---|---|
| $R^2$ | 0.05769283 | 0.29638461 | 0.79896364 |

actual blood glucose level and the predicted level is significant. The decision tree model is more accurate compared with the SVM model. However, there is obvious variance between the actual and the predicted blood glucose level when predicting the extreme high or low blood glucose values. The Random Forest model has the best prediction performance among the three time series regression models as it has closely predicted the actual blood glucose levels. It has able to reduce the variance between the predicted and the actual blood glucose level for both high and low blood glucose values.

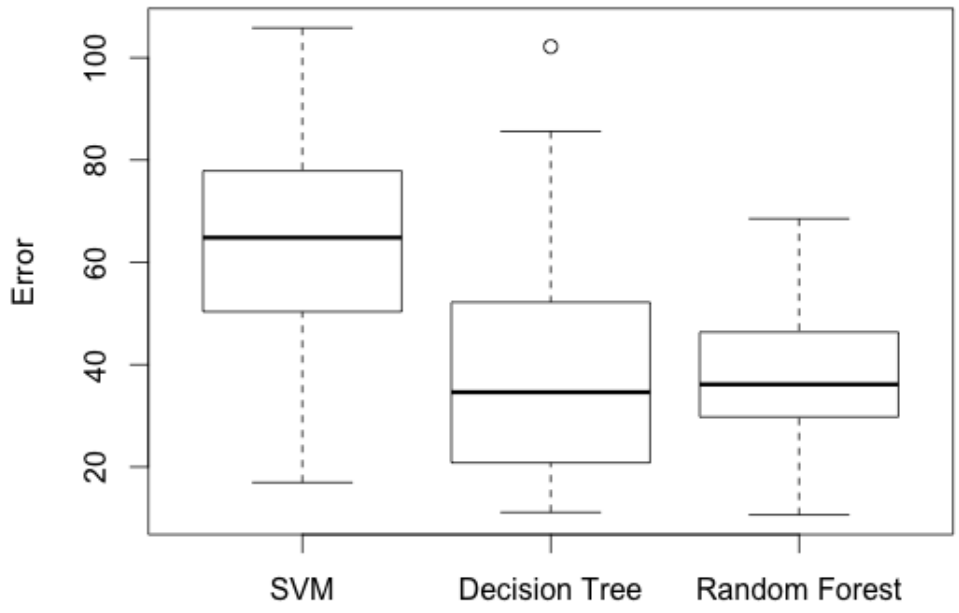


Figure 4.6. Boxplot of the root mean square of the three models

The table 4.1 and the box plot **??** summarize the root mean square error, mean absolute error and the determination of correlation of the three models. By comparing the root mean square error and mean absolute error values of the error table **??**, it is clear that both the decision tree and random forest models exceeds the SVM model by 39% and 42% respectively in terms of root mean square error and by 42% and 40% respectively in terms of mean absolute error. The reason behind the poor performance of the support vector regression lies in the pattern of the blood glucose levels. When observing the actual blood glucose pattern it is apparent that there are significant variations in the blood glucose level from morning to afternoon to evening to night. However the support vector regression tries to find a function which has less than $\varepsilon$ deviation from all the blood glucose values. However, the frequent and steep fluctuations of the blood glucose levels often result in data points with more than $\varepsilon$ deviation, making the model less accurate in predicting the blood glucose values.

Between the decision tree model and the random forest model, random forest model yields more accurate results in terms of root mean square error and it has a high coefficient of determination than the decision tree model. Further, random forest model's predictions tend to be closely distributed around the mean than the other two models. The reason behind it is that the random forest generates a large set of trees and uses the average prediction values among them.

### 4.1.3. Pooled Panel Model

In the second part of the experiment, we studied the impact of sample data size (i.e., frequency of blood glucose measurement) to prediction accuracy. Then we evaluated the performance of pooled panel data regression model as a remedy to the data sparsity problem of the individual regression mode. The time series regression prediction requires that the system maintains a large amount of historical data for the individuals whom we are going to make prediction for. However, as we have discussed, we may not have enough individual data to make predictions. People may not have time or condition to make frequent and regular blood glucose measurement.

We prepared 2 sub datasets from a randomly selected patient by removing the records in following order.

1. I0 : None of the records were removed
2. I1 : Every other record was removed which resulted in making the dataset half.

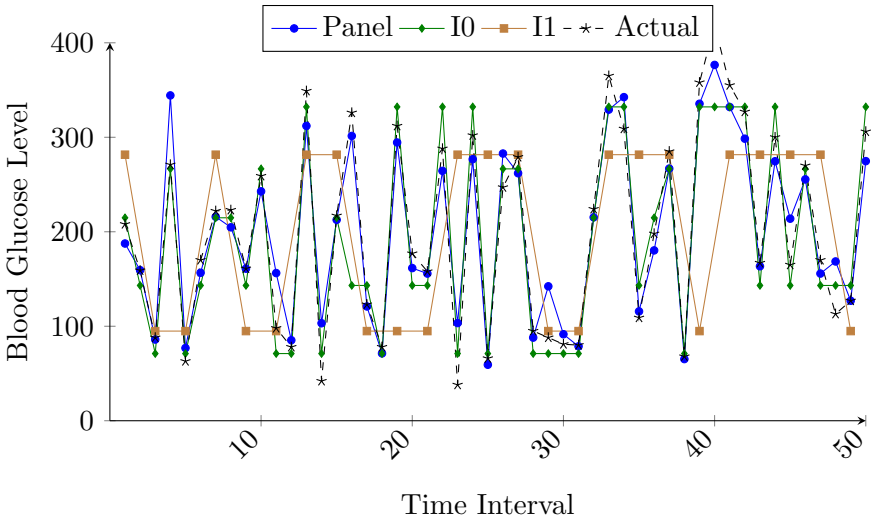


Figure 4.7. Predicted blood glucose levels using pooled panel data regression and different sized personal data (I0: individual-based prediction using the whole samples in the dataset, I1: Individual-based prediction with half of the sample size.)

Table 4.3. Comparison of Pooled panel data-based Prediction with Pre-Cluster based Predictions

| Metric | Pooled data | 2-Cluster | 3-Cluster | 5-Cluster | 9-Cluster | 43-Cluster |
|---|---|---|---|---|---|---|
| RMSE | 39.438 | 36.695 | 33.573 | 33.074 | 31.193 | 27.453 |

Both the dataset I0 and I1 were used by the random forest model to predict blood glucose levels. We the substituted the randomly selected patient's data with I1 dataset in the original dataset to use it in pooled panel data regression model. Figure 4.7 illustrates the prediction performance under different sized individual data. In the figure I0 represents individual-based prediction using the whole samples in the dataset; I1 represents individual-based prediction with half of the sample size. As shown in the figure, as the frequency of an individual's blood glucose measurements reduces, the prediction accuracy decreases. However, still the pooled panel regression model was able to predict fairly accurate results in a situation of missing data.

### 4.1.4. Pre-Clustered Panel Model

The third experiment will be based on clustering the dataset and applying the pooled panel regression. Table 4.3 compares the prediction performance of pooled panel data-based prediction with personalized pre-cluster-based prediction with different cluster size. Different cluster sizes were achieved by horizontally slicing the dendogram at different distances.

As can be seen from the table 4.3, personalized pre-clustered-based prediction is more accurate than the panel data regression without clustering. The root mean square error has been

Table 4.4. $R^2$ Comparison of Pooled panel data-based Prediction with Pre-Cluster based Predictions

| Metric | Pooled data | 2-Cluster | 3-Cluster | 5-Cluster | 9-Cluster | 43-Cluster |
|---|---|---|---|---|---|---|
| $R^2$ | 0.7681 | 0.8047 | 0.8014 | 0.8032 | 0.8156 | 0.8883 |

reduced when number of clusters are increased (number of members in a cluster shrinks). Further the reduced sized clusters have obtained increased $R^2$ value which mean their fit to the model has increased with the reduced cluster size 4.4. Main reason would have been that the similarity of the blood glucose levels increase with the cluster size decrease. Hence the negative impact of non similar patients are minimized.

In summary, the experimental results demonstrate that the proposed pre-cluster-based prediction can improve the accuracy of prediction compared with the pooled panel data-based prediction. On the other hand, it can also remedy the sparsity problem of individual data.

# 5. CHAPTER FIVE

## 5.1. Conclusion

As for the first part of the experiment, we have compared three machine learning techniques in predicting blood glucose level of an individual. The results shows that the random forest model with 68 trees yields the least error among the three models. This finding will be important when there is only one patient data is available, which makes it impossible to use panel data regression models. In addition to it, individual model has to be used when there is only one element remains after clustering a pool of data. The second conclusion of the experiment is that the pooled panel data model is an useful model when there are missing data in patients' records. The individual regression models lack accurate predicting of the blood glucose level where as the pooled panel data model displayed high accuracy as it used all the patient data in deriving the regression line. The third experiment was on applying clustering to the dataset in order to segment patients with similar qualities. The results showed that the accuracy and the fitness to the model increase as the cluster size gets smaller. It is difficult to arrive on an optimal cluster size where the accuracy and the fitness to the model would be smallest for all the datasets, it depends on the dataset as well.

## 5.2. Future Work

The blood glucose prediction is an area with ample of opportunity to work on and improve the existing models. Advancements to the existing regression models can be proposed in future. New regression models can be innovated in future. Regarding the experiment one, where we compared three individual based models, such new models can be applied and compare their performance with the existing techniques. In addition to testing new regression models, if a high order differential function can be derived to estimate the rate of change in human blood glucose levels, the blood glucose prediction can be modeled as an initial value problem. Methods such as Euler approximation, 2nd order and 4th order Runge Kutta methods will be able to use in predicting blood glucose values. One of the main advantage of such methods is one would not have to maintain a history of his blood glucose measurements. The proposed methods in the thesis rely on historic data and would fail to predict blood glucose level if a new patient arrives who does not has blood glucose values measured previously. A dataset with more features can be used in evaluating

the models' performances. New features can be of patients' age, sex, marital status, occupation ethnic group or type of food consumed. With regarding the clustering the dataset for the third experiment, one can try other pairwise distance functions to calculate the pairwise distance matrix to perform a better hierarchical clustering. In addition to it, once a dataset with new features and more patient instances are available, different clustering techniques can be performed to identify patients with similar properties.

# REFERENCES


[1] American Diabetes Association, "National Diabetes Statistics Report , 2014 Estimates of Diabetes and Its Burden in the Epidemiologic estimation methods," *National Diabetes Statistics Report*, pp. 2009–2012, 2014.

[2] D. Basak, S. Pal, and D. C. Patranabis, "Support Vector Regression," *Neuronal Information Processing - Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.

[3] M. S. Kirkman, V. J. Briscoe, N. Clark, H. Florez, L. B. Haas, J. B. Halter, E. S. Huang, M. T. Korytkowski, M. N. Munshi, P. S. Odegard, R. E. Pratley, and C. S. Swift, "Diabetes in older adults," *Diabetes Care*, vol. 35, no. 12, pp. 2650–2664, 2012.

[4] E. Benjamin, "Self-Monitoring of Blood Glucose: The Basics," *Clinical Diabetes*, vol. 20, no. 1, pp. 45–47, 2002.

[5] D. C. Klonoff, "The artificial pancreas: how sweet engineering will solve bitter problems." *Journal of diabetes science and technology*, vol. 1, no. 1, pp. 72–81, 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2769610{&}tool=pmcentrez{&}rendertype=abstract

[6] V. W. Bolie, "Coefficients of normal blood glucose regulation." vol. 16, pp. 783–788, 1961.

[7] G. Toffolo, R. N. Bergman, D. T. Finegood, C. R. Bowden, and C. Cobelli, "Quantitative estimation of beta cell sensitivity to glucose in the intact organism: a minimal model of insulin kinetics in the dog," *Diabetes*, vol. 29, no. 12, pp. 979–990, 1980.

[8] M. Derouich and a. Boutayeb, "The effect of physical exercise on the dynamics of glucose and insulin," vol. 35, pp. 911–917, 2002.

[9] M. S. Shanker, "Using neural networks to predict the onset of diabetes mellitus," *Journal of Chemical Information & Computer Sciences*, vol. 36, no. 1, pp. 35–41, 1996. [Online]. Available: http://linksource.ebsco.com/linking.aspx?sid=OVID:medline{&}id=pmid:8576289{&}id=doi:{&}issn=0095-2338{&}isbn={&}volume=36{&}

issue=1{&}spage=35{&}date=1996{&}title=Journal+of+Chemical+Information+{&}+Computer+Sciences{&}atitle=Using+neural+networks+to+predict+the+onset+of+diab

[10] C. Zecchin, A. Facchinetti, G. Sparacino, and C. Cobelli, "Reduction of number and duration of hypoglycemic events by glucose prediction methods: a proof-of-concept in silico study," *Diabetes technology & therapeutics*, vol. 15, no. 1, pp. 66–77, 2013.

[11] C. Marling, R. Bunescu, J. Shubrook, and F. Schwartz, "System Overview : The 4 Diabetes Support System," *Workshop Proceedings of the Twentieth International Conference on Case-Based Reasoning*, pp. 81–86, 2012.

[12] W. A. Sandham, D. J. Hamilton, A. Japp, and K. Patterson, "Neural network and neuro-fuzzy systems for improving diabetes therapy," in *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, vol. 20, no. 3, 1998, pp. 1438–1441.

[13] A. K. El-Jabali, "Neural network modeling and control of type 1 diabetes mellitus," *Bioprocess and biosystems engineering*, vol. 27, no. 2, pp. 75–79, 2005.

[14] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz, "A machine learning approach to predicting blood glucose levels for diabetes management," *Modern Artificial Intelligence for Health Analytics. Papers from the AAAI-14*, 2014.

[15] F. Pociot, A. E. Karlsen, C. B. Pedersen, M. Aalund, J. Nerup, E. C. for IDDM Genome Studies, and Others, "Novel analytical methods applied to type 1 diabetes genome-scan data," *The American Journal of Human Genetics*, vol. 74, no. 4, pp. 647–660, 2004.

[16] J. Han, J. C. Rodriguez, and M. Beheshti, "Diabetes data analysis and prediction model discovery using rapidminer," in *Future Generation Communication and Networking, 2008. FGCN'08. Second International Conference on*, vol. 3. IEEE, 2008, pp. 96–99.

[17] B. Sudharsan, M. Peeples, and M. Shomali, "Hypoglycemia Prediction Using Machine Learning Models for Patients With Type 2 Diabetes," *Journal of Diabetes Science and Technology*, vol. 9, no. 1, pp. 86–90, 2014. [Online]. Available: http://dst.sagepub.com/lookup/doi/10.1177/1932296814554260

[18] Saedsayad. Decision Tree Regression. [Online]. Available: http://www.saedsayad.com/decision{_}tree{_}reg.htm

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://link.springer.com/article/10.1023/A:1010933404324

[20] J. J. J. Groen, "Exchange Rate Predictability and Monetary Fundamentals in a Small Multi-Country Panel," *Journal of Money, Credit, and Banking*, vol. 37, no. 3, pp. 495–516, 2005. [Online]. Available: http://muse.jhu.edu/content/crossref/journals/journal{_}of{_}money{_}credit{_}and{_}banking/v037/37.3groen.pdf

[21] N. C. Mark and D. Sul, "Nominal exchange rates and monetary fundamentals. Evidence from a small post-Bretton woods panel," *Journal of International Economics*, vol. 53, no. 1, pp. 29–52, 2001.

[22] V. Cerra and S. C. Saxena, "The monetary model strikes back: Evidence from the world," *Journal of International Economics*, vol. 81, no. 2, pp. 184–196, 2010.

[23] Wikipedia. (2015) Mean absolute error — Wikipedia{,} The Free Encyclopedia. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Mean{_}absolute{_}error{&}oldid=660816639

[24] ——. (2016) Root-mean-square deviation — Wikipedia{,} The Free Encyclopedia. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Root-mean-square{_}deviation{&}oldid=712175603

[25] ——. (2016) Coefficient of determination — Wikipedia{,} The Free Encyclopedia. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Coefficient{_}of{_}determination{&}oldid=710897825