DIABETES DOSE TITRATION IDENTIFICATION MODEL

RATCHANEE KAEWTHAI

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(INFORMATION TECHNOLOGY MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2015

Thesis
entitled
DIABETES DOSE TITRATION IDENTIFICATION MODEL

.....
Miss Ratchanee Kaewthai
Candidate

.....
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Major advisor

.....
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer
Engineering)
Co-advisor

.....
Lect. Taweesak Samanchuen,
Ph.D. (Electrical Engineering)
Co-advisor

.....
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

.....
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer
Engineering)
Program Director
Master of Science Program in
Information Technology Management
Faculty of Engineering
Mahidol University

Thesis
entitled
DIABETES DOSE TITRATION IDENTIFICATION MODEL

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Information Technology Management)
on
December 24, 2015

.....
Miss. Ratchanee Kaewthai
Candidate

.....
Asst. Prof. Nantawat Sitdhiraksa,
M.D.,Ph.D.(Psychiatry)
Chair

.....
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Member

.....
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Member

.....
Lect. Taweesak Samanchuen,
Ph.D. (Electrical Engineering)
Member

.....
Prof. Somkiat Wattanasirichaigoon,
M.D., FRCST (General Surgery)
Member

.....
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

.....
Asst.Prof. Jackrit Suthakorn,
Ph.D. (Robotics)
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

This research unable to success without advices, supports and encouragements from many people. I would like to deeply grateful to my thesis advisor, Dr. Sotarat Thammaboosadee, co-advisor, Asst. Prof. Supaporn Kiattisin and Dr.Taweesak Samanchuen for invaluable help, constant encouragement and kindness in writing format proofing throughout the course of this research which not only the research methodologies but also many other methodologies all time ago.

Special thanks to my family, colleague and my friends to suggest and all their support throughout the period of this thesis. Additionally, Those whose names are not considered here but have greatly encouraged and inspired us until complete research.

Ratchanee Kaewthai

DIABETES DOSE TITRATION IDENTIFICATION MODEL

RATCHANEE KAEWTHAI 5737290 EGIT/M

M.Sc. (INFORMATION TECHNOLOGY MANAGEMENT)

THESIS ADVISORY COMMITTEE : SOTARAT THAMMABOOSADEE, Ph.D., SUPAPORN KIATTISIN, Ph.D., TAWEESAK SAMANCHUEN, Ph.D.

ABSTRACT

Diabetes is a chronic disease that requires continuous treatment throughout lifespan and it risks developing a number of serious health problems, leading to high treatment cost. Admitted diabetes inpatients should receive the appropriate treatment in order to reduce severe complications and premature death. This research aims to develop the classification model for diabetic medication adjustment based on historical medical record of diabetic inpatients by applying three algorithms; Decision Tree, Naïve Bayes, and Artificial neural network (ANN). By comparison of the results of each method, Decision Tree outperformed the others for Independent Dose Titration dataset (IDT). On the other hand, Artificial Neural Network algorithm could generate a model to adjust medication adjustment with high accuracy and ROC Curve for Historical Dose Titration dataset (HDT). An additional enhancement of ANN tuning was also experimented with HDT dataset and insulin. The results of this paper could support the decision making in medication adjustment of diabetes inpatients, particularly those with type-2 diabetes.

KEY WORDS: DOSE TITRATION / DIABETES / DATA MINING
IDENTIFICATION MODEL / MEDICAL MANAGEMENT.

49 pages

แบบจำลองสำหรับระบุการปรับยาในโรคเบาหวาน
DIABETES DOSE TITRATION IDENTIFICATION MODEL

รัชนี แก้วไทย 5737290 EGIT/M

วท.ม. (การจัดการเทคโนโลยีสารสนเทศ)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ : โษฑศ์รัตต ธรรมบุษดี, Ph.D., สุภาภรณ์ เกียรติสิน, Ph.D., ทวีศักดิ์ สมานชื่น, Ph.D.

บทคัดย่อ

โรคเบาหวานเป็นโรคเรื้อรังที่ต้องการการดูแลรักษาตลอดชีวิตและเพิ่มความเสี่ยงต่อการเกิดโรคแทรกซ้อนซึ่งเป็นอันตรายและมีค่าใช้จ่ายสูง ผู้ป่วยที่เข้ารับการรักษาในโรงพยาบาลควรจะได้รับการดูแลรักษาอย่างเหมาะสมเพื่อลดอัตราการเกิดภาวะโรคแทรกซ้อนและการเสียชีวิตก่อนวัยอันควร งานวิจัยชิ้นนี้มีวัตถุประสงค์ที่จะพัฒนาแบบจำลองสำหรับระบุการปรับยาในโรคเบาหวาน โดยใช้ประวัติการรักษาของผู้ป่วยเป็นข้อมูลพื้นฐานมาประยุกต์ร่วมกับเทคนิคการทำเหมืองข้อมูล 3 อัลกอริธึม ได้แก่ ต้นไม้ตัดสินใจ, นาอีฟ เบย์, โครงข่ายประสาทเทียม จากผลการทดลองพบว่าวิธีต้นไม้ตัดสินใจเมื่อใช้กับข้อมูลชุดไอดีที (IDT: Independence Dose Titration dataset) แสดงแบบจำลองสำหรับระบุการปรับยาได้แม่นยำและถูกต้องสูงกว่าอีก 2 วิธีและวิธีโครงข่ายประสาทเทียม สามารถสร้างแบบจำลองสำหรับระบุการปรับยาได้ถูกต้องและแม่นยำสูงกว่าวิธีอื่นเมื่อใช้กับข้อมูลชุดเอชดีที (HDT: Historical Dose Titration dataset) และหากมีการปรับค่าพารามิเตอร์ก็ส่งผลให้ประสิทธิภาพของโมเดลสูงขึ้น โดยได้ทำการทดลองกับข้อมูลยาอินซูลิน (Insulin dataset) ซึ่งจากผลการทดลองสามารถสร้างแบบจำลองสำหรับระบุการปรับยาในโรคเบาหวานโดยเฉพาะผู้ป่วยโรคเบาหวานประเภท 2 ที่ช่วยสนับสนุนการตัดสินใจของแพทย์ตลอดจนวางแผนการรักษาที่เหมาะสมกับผู้ป่วยเบาหวานแต่ละรายโดยใช้วิธีเหมืองข้อมูล

49 หน้า

CONTENTS

CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURES

CHAPTER I
INTRODUCTION

This research provides the intelligent data analysis in medical field approaching to a machine learning perspective based on the historical data of diabetic inpatients. The Classification model is proposed to identify the diabetes medication adjustment. The aim of the proposed model is to support the decision of the medical and treatment plan that suits the individual characteristics of type-2 diabetes inpatients.

1.1 Background and significance of the problems.

Chronic disease is becoming serious health problem of inability and death in worldwide [1]. Diabetes is a chronic disease that requires continuous treatment for a time and increased risk of developing to serious health problems, which mean high treatment cost and possible caused dangerous effects such as heart failure, stroke, kidney failure, eye damage and foot damage [2,3]. The current global diabetes population is approximately 387 million people and expected to increase to 205 million people with in year 2035 according to the rate of diabetes in USA. Nowadays, the total 19.50% of worldwide diabetic patients tends to increase with 27.25% of the cost of diabetes per person that is higher than the average rate of 188.62% [4].

Diabetes is caused by high glucose level of blood when the pancreas does not produce high enough insulin to handle the storage of sugar from the blood into the cells in order to maintain the blood sugar levels to the normal state. This causes the patient to be severe hyperglycemia. Diabetes is a chronic and incurable disease. Diabetic patients can only control sugar levels in the normal range by dieting, exercise and medication which they would live as the normal life like others. Medically, there are two important types of medication which are insulin and oral questions [5].

Currently, there is no research representing that which medication is the best for reducing blood glucose level. Therefore, the selection of drug usage to lower

blood sugar level will depend on the effectiveness of reducing blood sugar level to meet the target, the effectiveness against the complications and adverse reaction against drug usage, the patient's tolerance against drug including the appropriate expenses [6]. Thus, it is interesting to analyze the history of diabetes inpatient for building up the design model that suits the individual characteristics of type-2 diabetes inpatients in the future by data mining.

1.2 Research Objectives

This research is aimed to develop the Classification model for diabetes medication adjustment based on historical medical record of diabetic patients. The classification model is driven by the data mining approaches.

1.3 Research Scopes

Data in this cross sectional research taken from Health Facts database (Cerner Corporation, Kansas City, MO) [7], a national data warehouse of hospitals in the United States which collects comprehensive clinical records across hospitals throughout country. Data was an extract representing 10 years (1999 - 2008) of type-2 diabetes noncritical inpatients at 130 hospitals integrated delivery networks throughout the United States.

1.4 Expected Results

The expected result of this research is the diabetic medication adjustment model to support the medical decisionthat suits the individual characteristics of diabetes patients.

The next chapter describes the diabetes, the medication, data mining, and related research. Then, the chapter III explains how to operate the research. Chapter IV presents the experimental results. Finally, conclusion and future works are described.

CHAPTER II
LITERATURE REVIEW

This research start from the gathering data of diabetes inpatients, and study the method for building the pattern of diabetes dose titration. This chapter discusses the concepts of research study as follows:

2.1 Diabetes,
2.2 Pharmacological Therapy,
2.3 Dose Titration for diabetes inpatients,
2.4 Data Mining,
2.5 Classification Tasks,
2.6 Model Evaluation,
2.7 Related Research.

2.1 Diabetes

Diabetes is a chronic, non-communicable disease group. This is a condition when the glucose in the blood reaches too high level. Normally, glucose levels in the blood are controlled by the hormone insulin that is generated by the pancreas. Diabetes are caused by the severe hyperglycemia when the pancreas does not produce enough insulin or the body's cells do not react with insulin and make the condition with the high level of glucose accumulated in the blood. Diabetes leads to high risks in being heart disease, stroke, kidney failure, blindness, and lower limb amputation [5,6].

Type-2 diabetes is formerly known as "noninsulin-dependent diabetes" or "adult onset diabetes". Type-2 diabetes is caused from inadequate insulin production or non-respond of insulin in the body that making too high levels of glucose in the blood. Type-2 diabetes patients are higher than other types of diabetes (approximately 90-95% of worldwide diabetes rate). Type-2 diabetes can lead to severe health

complications such as heart and blood vessel disease, nerve damage (neuropathy), kidney damage (nephropathy), eye damage, foot damage, hearing impairment, skin conditions, and Alzheimer's disease.

Diagnosis of Diabetes are considered by the criteria of Hemoglobin A1C (A1C) or plasma glucose, either the fasting plasma glucose (FPG) or the 2-h plasma glucose (2-h PG) value after a 75-g oral glucose tolerance test (OGTT) [8,9,10]. Type-2 diabetes is detected by testing with the prospects aged start from 45 years old who are overweight. If the testing results are normal, repeated testing should be done again within 3 years [3].

2.2 Pharmacological Therapy

Therapy of Type-2 diabetes can be done by regulating the blood sugar levels with both exercise and diet. However most of diabetes patients do not achieve by mentioned therapy. Thus insulin therapy or medications are considered as the proper therapy concomitant with the doctor’s prescription. The prescription rely on many variables such as blood sugar level and other health problems of the diabetes patients. The appropriate therapy of each patient may varies by any factors which is considered and decided by the doctors [8].

Type-2 diabetes treatment examples [8]:

- **Metformin** is considered as the first drug prescribed to Type-2 diabetes patients who are failure in diet. This drug works by deduction of sugar absorbance and prevention of glucose production from pancreas that make the effective insulin usage in the body.

Sulfonylureas are the medicines use to help the body to emit more insulin, but may be possibly cause to gain weight and low blood sugar as the side effects. Examples of medicines in this group are Glyburide, Glipizide, and Glimepiride.

Meglitinides are the medicines which have the same qualification like Sulfonylureas but act faster and have a lower risk for making low blood sugar than Sulfonylureas do. Although they don’t stay active in the body for so long, but still

cause for weight gain. Examples of medicines in this group are Repaglinide and Nateglinide.

Thiazolidinediones are the medicines work like Metformin in making the body's tissues more sensitive to insulin and weight gain. These medications do not considered as the first priority to treat diabetes patients because they give the serious side effects such as increasing risk of heart failure and fractures. Rosiglitazone and Pioglitazone are the examples of medicines in this group.

DPP-4 inhibitors are the medicines use to reduce blood sugar levels. Seemingly they don't cause to gain weight and have a modest effect. Examples of medicines in this group are Sitagliptin, Saxagliptin, and Linagliptin.

GLP-1 receptor agonists are the medicines assisting to lower blood glucose levels. Although these medications slow digestion, but these are not recommended for use alone. This medication may be cause to incur side effects such as nausea and an increased risk of pancreatitis. Exenatide and Liraglutide are the examples of the medicines in this class.

SGLT2 inhibitors are the newest diabetes medicines that use for preventing the kidneys from reabsorbing glucose in the blood. Instead, the glucose is eliminated in the urine.

Insulin therapy is suited and often use for some type-2 diabetes patients. It will be considered as the first priority in diabetes treatment because of its benefits.

2.3 Dose Titration for diabetes inpatients.

Diabetes inpatients are classified as two groups, Critically ill patients and Non-critically ill patients [8]. All inpatients must be explicitly stated their diabetes type in medical record when admit in the hospital and diabetes arrangement instructions should be supplied.

Critically ill patients will be initially treated by insulin infusion when the hyperglycemia condition is greater than 180 mg/dl(10 mmol/L) in order to control glucose level within the range of 140–180 mg/dl (7.8–10 mmol/L). For some selected patients, may be strictly glucose level within the range of 110–140 mg/dL (6.1–7.8

mmol/L). The patients in this type require an intravenous insulin protocol to effectively regulate glucose range and reduce risk for severe hypoglycemia [8].

Non-critically Ill patients will be initially treated by insulin to control pre-meal glucose level within the range of <140 mg/dl and after-meal glucose level within the range of <180 mg/dl. The stringent glycemic control may be suited for regulate glucose within normal level, and the less stringent glycemic control may be suited for the patients with severe comorbidities. For the poor oral intake patients prefer to treat with a basal plus correction insulin regimen. The diabetes inpatients will be considered to obtain A1C test if no fasting plasma glucose testing results in the prior 3 months and have the hyperglycemia condition in the hospital but no asymptomatic. The hyperglycemic patients should have properly monitor and closely follow up testing.

2.4 Data Mining

Data mining [11] is the knowledge discovery method that mines from large amounts of data, including databases, data warehouses, the web, other information repositories, or data that are dynamically streamed into the organization's system.

The CRISP-DM Process

Cross-Industry Standard Process [12] for Data Mining (CRISP-DM) was developed in 1996 to adapt data mining to general business strategy. This process consists of 6 phrases lifecycle of the data mining project is shown in Figure 2.1.

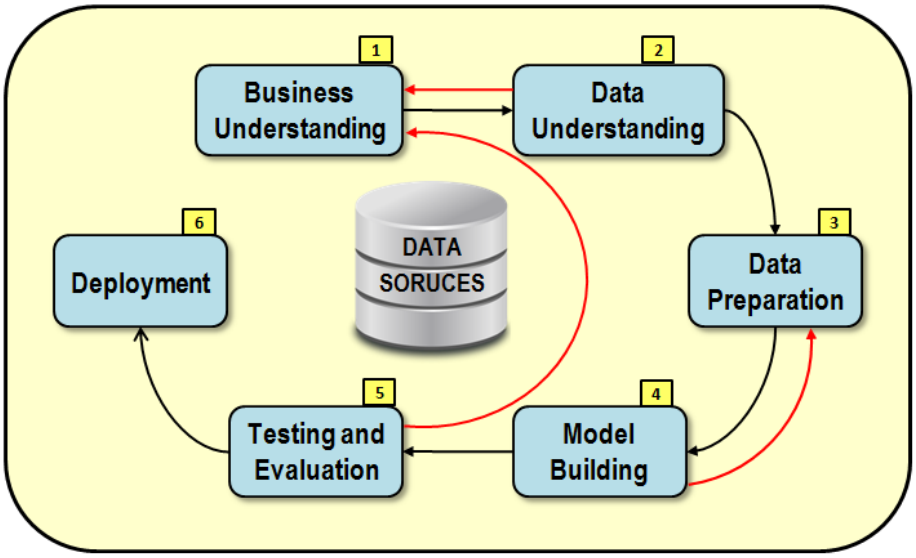


Figure 2.1 Data mining process.

1) Business understanding phase

The first phase is defining business requirements and objectives, translate into data mining problem definition, and prepare the initial tactic to achieve the objectives.

2) Data understanding phase

This phase is performed by accumulating the data, operate the exploratory data analysis (EDA) in order to familiarize with data and assess data quality.

3) Data preparation phase

During this phase, raw data is prepared, selected, and transformed into structured data set that suited for modeling.

4) Model building phase

This phase is selecting and applying the proper modeling techniques, calibrate and adjust model to optimize the results. In case of the additional data may be required, looping back to data preparation phase if necessary.

5) Testing and Evaluation phase

The model generated in the modeling phase is evaluated to ensure the quality and effectiveness before deployment. The objectives is determined the achievement, and making decision in related to data mining results.

6) Deployment phase

The last phase is applying live created model within an organization's decision making processes. The deployment phase can be done as simple or complex implementation, it depends on business requirement of each organization.

2.5 Classification Tasks

Classification is one of basic data mining technique using to classify item in a data set into one of predetermined class or group.

Model building is the first step in the classification algorithms. Data divided to training data and testing data. Some of classification technique such as decision tree, Artificial Neural Network, and Naïve Bayes makes use of mathematical techniques.

In the classification, training dataset is bring to imported to process for learning system. Training data include historical data such as common attribute and attribute classthat is interested in the target. Also important to consider is the attribute data will be correlated with the class attribute is the answer, it will create a model that is efficient and very reliable.

After modeling complete. Next step is model evaluation for test the validity of the model. The measure commonly used any method such as Precision, Recall, F-measure and accuracy by measuring data and Training set and Testing set.

After model evaluation is complete after measuring the performance of the model, the results are satisfactory, then taking the lead to build a model used to predict new data.

2.5.1 Decision Tree

Decision Tree [13] is one of supervised learning algorithm that widely used in identification classification model. This model is represented by tree diagram structure. Leaves represent class labels and branches represent correlation of attributes leading to each class labels. By tree based representation, the decision tree could

explicitly interpret, describe the results and enable user to analyze data and make decision more precisely.

The basic concept of decision tree algorithm is shown in Figure 2.2.

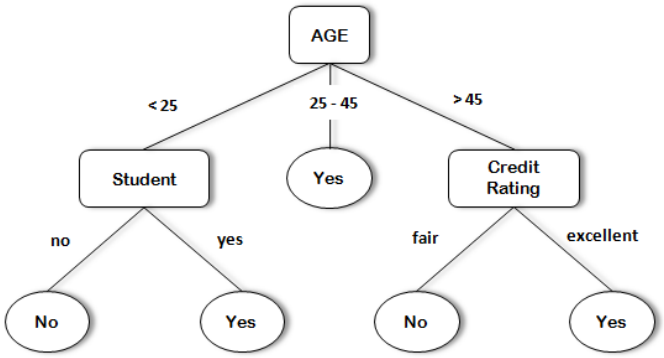
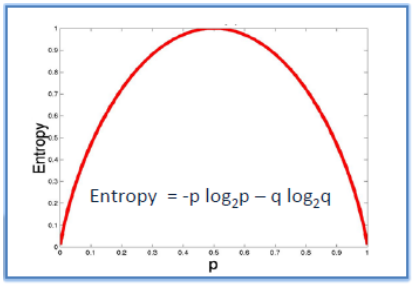


Figure 2.2 Decision tree concepts.

The core algorithm of decision tree is building up classification model in a tree structure form which comprise of a root node, branches, and leaf nodes. Dataset will be broken down into smaller subsets. Decision trees can perform both nominal and numerical data. Leaf node represents a decision. Root node is the topmost decision node in a tree which represents to the best predictor.

J. Ross Quinlan developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). ID3 algorithm uses entropy to calculate the homogeneity of a sample. The trees are constructed in a top-down recursive divide-and-conquer manner. If the entropy equal to zero , it means the sample is completely homogeneous. And if the entropy is one, the sample is equally divided. The Entropy equation is shown in Figure 2.3.

C4.5 [14] was a well-known decision tree technique developed from ID3 by Ross Quinlan in 1993. The main concept of this method based on information gain same as ID3. The advantage of C4.5 is ability to handle both continuous data and discrete data, including missing data. And the created decision tree could be pruning without impact on the accuracy.



Entropy = -0.5 $\log_2$0.5 – 0.5 $\log_2$0.5 = 1

Figure 2.3 Entropy equation [15].

The process of constructing a decision tree is finding attribute given the highest information gain as follow:

Firstly, the target's entropy is calculated. After that, the dataset is separated on the different attributes. The entropy of each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy. Next step, the attribute which have the largest information gain is selected as the decision node. The information gain is based on the decrease in entropy after a dataset is split on an attribute. Then a branch with entropy of 0 is a leaf node, and a branch with entropy more than 0 needs further splitting. Finally, the ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

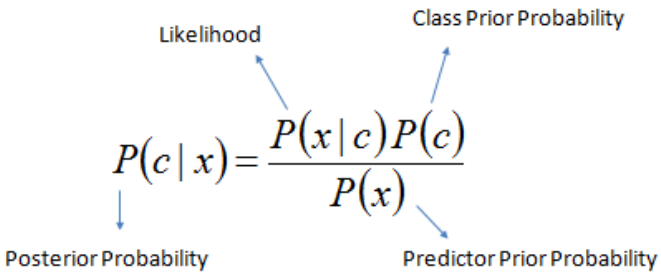
By mapping the root node to the leaf nodes one by one, a set of decision rules are generated.

2.5.2 Naive Bayesian

Naive Bayesian [13] is one of a simple popular classifier algorithm based on probability. The core concept of this method is to search for the probability of the formerly unseen instance of each class, and then select the probable class. Naïve Bayesian classifiers are easy to implement, not sensitive to irrelevant features and widely used in making decisions about treatment processes.

The Naive Bayesian classifier is the statistical classifier based on Bayes' theorem by finding out calculation method of the posterior probability., P(c|x), from P(c), P(x), and P(x|c).

Naive Bayes classifier was assumed that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence. The Bayes' theorem equation is shown in Figure 2.4.



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 2.4 Bayes' theorem equation. [11].

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

The posterior probability can be performed by generating the frequency table for each attribute against the target, and then transforming to the likelihood tables and calculating the posterior probability for each class by using the Naïve Bayesian equation. The prediction result is the highest posterior probability class.

2.5.3 Artificial Neural Network

An artificial neural network (ANN) is a computational model that simulated the biological neurons such as human brain. The natural neuron received the strong signals through synapses in order to activation and then sending a signal to the axon and other synapses

An ANN is comprised of three types of neurons including input nodes, hidden nodes, and output nodes. These basically are multiplied by weights and then processed by a mathematical method which determines the activation of the neuron. The strength of their connections to one another is assigned a value based on their strength: inhibition (maximum being -1.0) or excitation (maximum being +1.0). If the value of the connection is high, then it indicates that there is a strong connection.

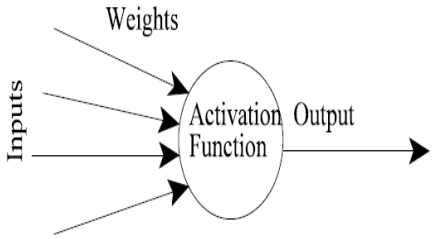


Figure 2.5 An artificial neural. [11].

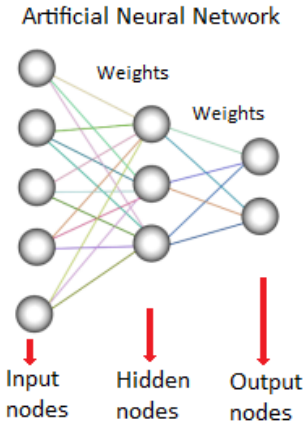


Figure 2.6 Artificial neural network. [11].

There are two basic types of neural networks: feed-forward and feed-back networks.

A feed-forward network consists of inputs, outputs, and hidden layer. This is a non-repeated network because the signals are transmitted by only one direction. Input data is go through a layer of processing elements and then calculated by based on a weighted sum of each input. The input values of next layer will be feed forward by using prior calculating results. This cycle will be repeated through all the layers and finally determines the output. Sometimes a threshold transfer function is used to quantify the output.

A feed-back network a network that the signals are transmitted in both directions using loops. This method is a non-linear dynamic system. The connection between neurons are allowable to find out the optimization or until it reaches the stability.

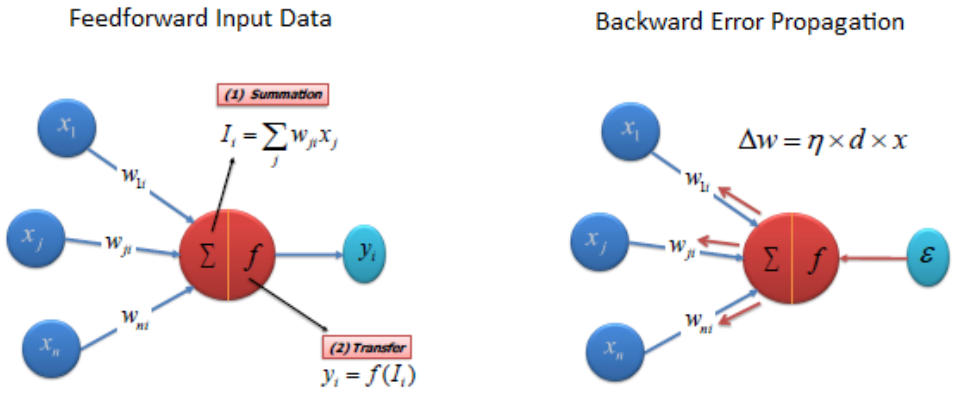


Figure 2.7 Feed-forward and Feed-back networks [11].

Activation Functions : Activation Functions: This function uses to transfer input signals to output signals. Unit step (threshold), sigmoid, piecewise linear, and Gaussian are the most commonly used transfer functions.

Unit step (threshold) : The output is set at one of two levels, depending on whether the total input is greater than or less than some threshold value.

Sigmoid : Logistic and tangential are types of this function which is differentiated by the range of values. The range of logistic function is from 0 and 1, while the range of tangential function is from -1 to +1.

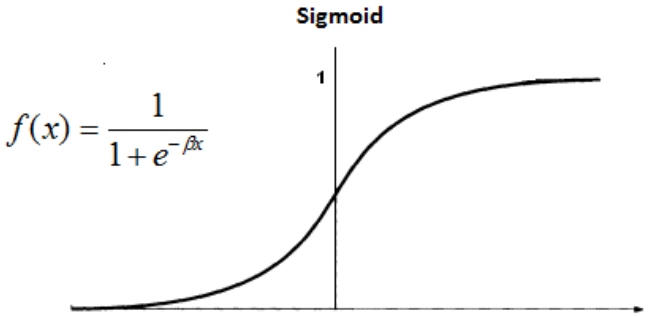


Figure 2.8 Example of Sigmoid function [13].

2.6 Model Evaluation

This research used 10-foldcross validation as a model evaluation method. K-Fold cross-validation [13] is the model evaluation method. The target data will be equally segmented into K folds and performed K times both training and testing. For example, in iteration of 10 folds, data will be partitioned in 10 subsets. The 1st partition is separated for testing and the remaining of 2-10 folds are using for train the model and make the rotation until reach 10 times. In classification technique, the estimated of accuracy is calculated by the summary of exact classifications derived from the K iterations, and then divided by the total number of tuples in data.

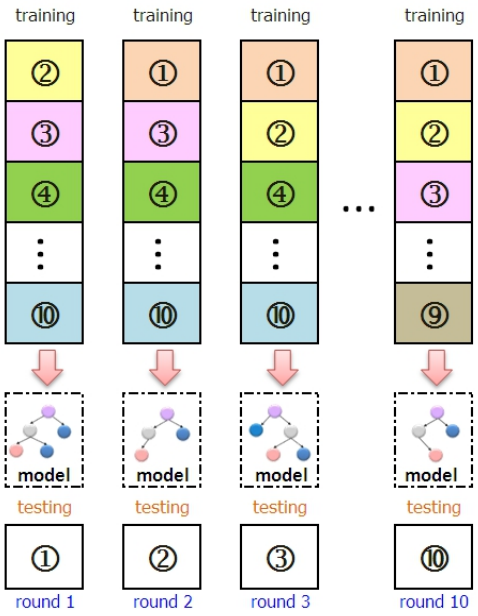


Figure 2.9 Example of 10-fold cross validation [13].

Data in each fold was consecutively trained and tested for model's performance evaluation. In each round was generated the different value of TP (true positive), TN (true negative), FN (false negative), and FP (false positive) to fill in confusion matrix. Finally the results were summed up and then calculated Precision, Recall, F-measure and Accuracy .

2.7 Related Research

Seksunti et al. [15] proposed research about Diabetes inpatient's knowledge discovered by Data Mining. Database of 12,000 chronic diabetes inpatient at Nongbua-Rawae Hospital during year 2006-2010 were used as a case study. Researchers analyzed patient's data by applying classification algorithms; Decision Tree with K-fold Cross Validation and Train-Test validation in mining knowledge Regarding to the model evaluation, the acquired experimental results were generated 12 rules with 83% of the accuracy. Nevertheless this work was proposed classification model in drug prescription using only patient's records. As a result the author propose

new methodology to create identification model for Diabetes titration by considering the individual drug dosage, lab results and historical therapy records as the database to build up model. With the combination of mentioned data, the acquired model is more accuracy and better support decision making in Dose Titration for type-2 diabetes inpatients.

Cook et al.[16] introduced the Intelligent Dosing System ; the mathematical model that using dose response data to calculate new dose and applied to Insulin adjustment for diabetes patients therapy.

Retrieved data was derived from an electronic Diabetes Patient Tracking System of a large urban outpatient diabetes clinic in between year 1991 - 2001. This work proposed method by based on mathematical model only, no related to Data Mining Techniques which is more effective, more complex data and provide better identification results for In-Hospital dose adjustment of Diabetes management.

However this research did not adopt data mining techniques in their Intelligent Dosing System (IDS). Moreover the input data was less complicate in drug's details and patient's details. Used data in this paper was comprised of 315 patients' record only. While Diabetes Dose Titration Identification Model retrieved input data over 100,000 records for processing that cause the results have more accuracy and more reliable than IDS.

Canadian Diabetes Association Clinical Practice Guidelines Expert Committee [17] was the Clinical Practice Guidelines for In-Hospital Management of Diabetes. This research guided that for in-hospital patients, blood sugar level's controlling is more significant than food controlling and Diabetes management. In-hospital Management of Diabetes is used as the Clinical Practice Guidelines and be categorized in 10 main topics covering all managerial aspects about in-hospital diabetes patients. Researcher proposed many theories for being guidelines in effective dose titration management.

This chapter reviews the related information and the previous researches. Each part contains the important information that gives the good idea for development. The main topic consists of the diabetes such as diabetes type, pharmacological therapy, medication management for diabetes inpatients and data mining. For the next chapter, the methodology of this research will be discussed.

CHAPTER III
METHODOLOGY

This chapter focuses on the use of classification techniques by Decision tree, Naïve Bayes, and ANN for the analysis of historical data of diabetes inpatients. The classification model is proposed to identify the diabetes medication adjustment. The aim of the proposed model is to support the decision of medical and treatment plan that suits the individual characteristics of type-2 diabetes inpatients.

3.1 Research Process

This research uses the classification techniques to identify the diabetes medication adjustment. There are various processes as shown in Figure 3.1.

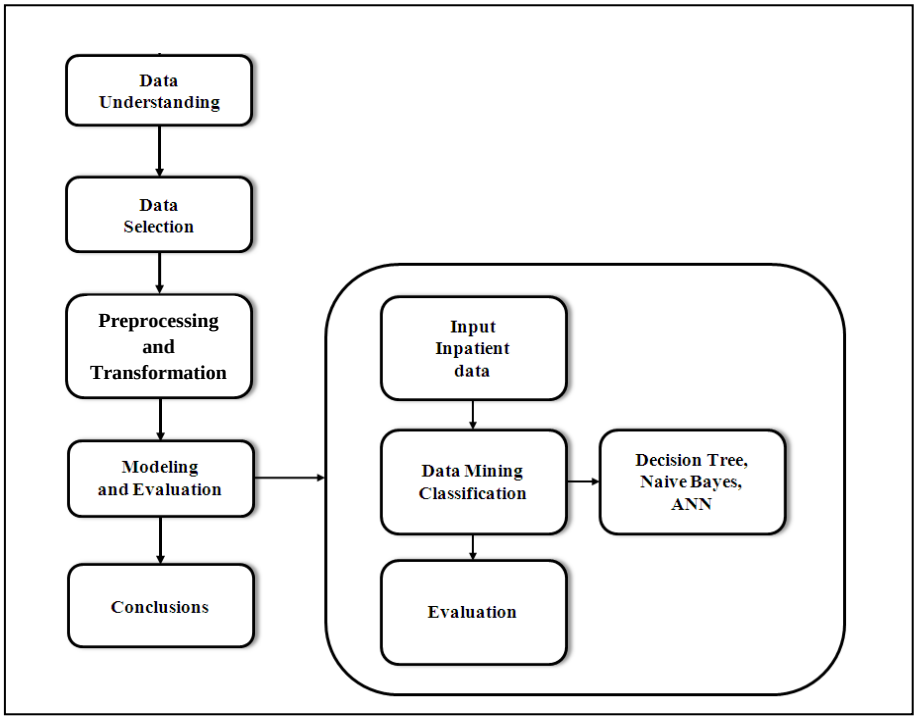


Figure 3.1 Research Process.

3.1.1 Data understanding

This research uses the Health Facts database from Cerner Corporation, Kansas City, MO, which is a national data warehouse that collects the comprehensive clinical records across hospitals throughout the United States [14]. The dataset consists of historical record which register 101,766 diabetes inpatient, 55 features including patient's medical record, lab results, and drug categories .

3.1.2 Data Selection

Next process, the raw data, consisting of patients data, lab results and drug type are selected accordingly.

Patients data are primarily filtered by the change of medications status as shown in positive and removed are the missing value feature and delete the irrelevant attributes which do not impact the model building. The remainder attributes are given as : Encounter ID, Patient number, Gender, Age, and Diagnosis_1.

Lab results are selected from Glucose serum test result and A1c test result. Drug types comprise of 24 features covering values : up, steady, down, and no. Then, the negative values (no) are deleted. After that, the remainder data are filtered by statistical method until receiving the final results of 8 features.

All of 7 features are Metformin, Glimepiride, Glipizide, Glyburide, Pioglitazone, Rosiglitazone, and Insulin.

The data selection for building the predict model consists of 15 attribute as shown in Table 3.1.

Table 3.1 The 15 final selected attributes.

Feature name	Description and values
Encounter ID	Unique identifier of an encounter
Patient number	Unique identifier of a patient
Gender	Values: male, female

Table 3.1 The 15 final selected attributes. (Cont).

Feature name	Description and values
Age	Grouped in 10-year intervals: (0, 10], (10, 20], …, (90, 100]
Diagnosis_1	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
Glucose serum test result	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
A1c_test_result	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.
7 features for medications	For the generic names: Metformin, Glimepiride, Glipizide, Glyburide, Pioglitazone, Rosiglitazone and Insulin the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter. “down” if the dosage was decreased. “steady” if the dosage did not change.

3.1.3 Data Preprocessing and Transformation

Initial raw data is preprocessed and transformed with some attributes in order to resize data and to reduce the processing time purpose. The complete selected data supplies in this research comprising of 15 attributes: Encounter ID, Patient number, Gender, Age, Diagnosis_1, Glucose serum test result, A1c_test_result, Metformin, Glimepiride, Glipizide, Glyburide, Pioglitazone, Rosiglitazone and Insulin.

Some attributes are preprocessed accordingly; Age, Diagnosis_1, Glucose serum_test_result, A1c_test_result, Metformin, Glimepiride, Glipizide, Glyburide, Pioglitazone, Rosiglitazone, and Insulin.

Age attribute is firstly deleted with the age under 40 years old, then it would transform the remained data by grouping into 6 groups consisting of 40(rang 40-50), 50(range 50-60), 60 (range 60-70), 70 (range 70-80), 80 (range 80-90), and 90 (range 90-100).

Diagnosis_1 attribute is categorized by the knowledge based on the research such as; The Impact of HbA2c Measurement on Hospital Readmission Rates : Analysis of 70,000 Clinical Database Patient Records [18], which suggests that the relationship between the probability of readmission and the HbA1c measurement depends on the primary diagnosis and the used database on the same work, as shown in Table 3.2.

Table 3.2 Diagnosis _1 grouping by ICD 9 codes.

Group name	ICD9 codes	Description
Circulatory	390–459, 785	Diseases of the circulatory system
Respiratory	460–519, 786	Diseases of the respiratory system
Digestive	520–579, 787	Diseases of the digestive system
Diabetes	250.xx	Diabetes mellitus
Injury	800–999	Injury and poisoning
Musculoskeleta	710–739	Diseases of the musculoskeletal system and connective tissue
Genitourinary	580–629, 788	Diseases of the genitourinary system
Neoplasms	140–239	Neoplasms
Other	780, 781, 784, 790–799	Other symptoms, signs, and ill-defied conditions
	240–279, without 250	Endocrine, nutritional, and metabolic diseases and immunity
	680–709, 782	disorders, without diabetes
	001–139	Diseases of the skin and subcutaneous tissue
	290–319	Infectious and parasitic diseases
	E–V	Mental disorders
	280–289	External causes of injury and supplemental classification
	320–359	Diseases of the blood and blood-forming organs
	630–679	Diseases of the nervous system
	360–389	Complications of pregnancy, childbirth, and the puerperium
	740–759	Diseases of the sense organs
		Congenital anomalies

Glucose_serum_test_result is transformed into 4 ranges based on the result value consist of 0 (“none”), 1 (“normal”), 2 (“>200”), and 3 (“>300”).

A1c_test_result is transformed into 4 ranges based on result value consist of 0 (“none”), 1 (“normal”), 2 (“>7”), and 3(“>8”).

Medical Compound of 7 drugs are classified into 4 classes : Metformin(Metformin), Sulfonylurea(Glimepiride, Glyburide, Glipizide), Thiazolidinedione (Pioglitazone, Rosiglitazone), and Insulin(Insulin) which are based on American Diabetes Association Standard of Medical Care in Diabetes 2015.

After finishing the primary preprocessing stage, data are classified into 2 datasets: Independent Dose Titration dataset (IDT) and Historical Dose Titration dataset (HDT) before performing Data Mining process.

Details of IDT dataset are shown in Table 3.3.

Table 3.3 Detail of IDT dataset.

Feature name	Description and values
Gender	male, female
Age	Range of patients age; 40, 50, 60, …, 90
Diagnosis 1	The primary diagnosis.
Glucose serum test result	Value : 0,1,2 and 3
A1c_test_result	Value : 0,1,2 and 3
4 medical classes	For the 4 medical classes : Metformin, Sulfonylurea, Thiazolidinedione and Insulin the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter. “down” if the dosage was decreased. “steady” if the dosage did not change.

IDT dataset was divided to 4 classes by drug's group. Due to the imbalance in retrieved datasets , caused each dataset must be resample before prior than. Each class comprise of data subset as per Figure 3.2.

Gender	Age	Diag. 1	Glucose	A1c	Sul
Female	50	Injury	1	0	Steady

Figure 3.2 Example of IDT subset data for Insulin modeling.

HDT dataset was similar the first one, but data was more complicate in details than IDT dataset. The historical record of dose titration was added by two previous times retrospectively. Because of the imbalance in retrieved datasets , caused each dataset must be resample prior than. Each class consist of data subset as per Figure 3.3.

Gender	Age	Diag. 1	Glucose	A1c	Sul_2	Sul_1	Sul
Female	50	Injury	1	0		Down	Steady

TZD_2	TZD_1	TZD	Ins_2	Ins_1	Ins	Met_2	Met_1	Met
	Up	Steady					Down	Down

Figure 3.3 Example of HDT subset data for Metformin modeling.

3.1.4 Modeling and evaluation using data mining

Model building.

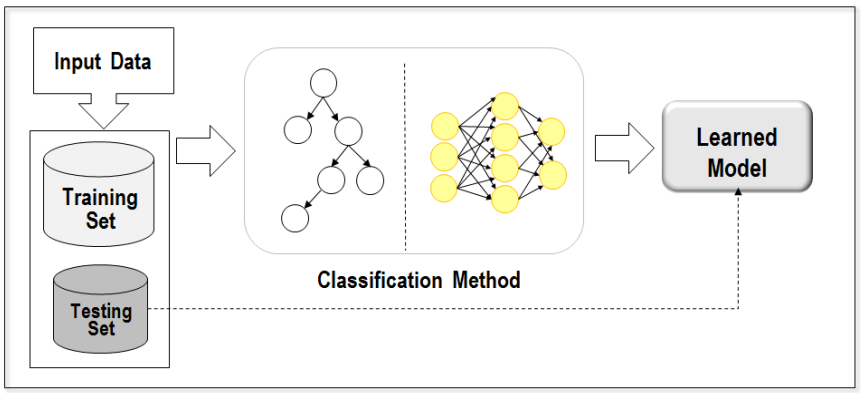


Figure 3.4 Classification Model.

As shown in Figure 3.4, this research proposes to generate the diabetes dose titration identification model by applying three classification algorithms; Naïve Bayes, Decision Tree, and Artificial Neural Network . The experiment was initiated by data selection. Then selected data is separated into two sets as IDT dataset and HDT dataset before performing Data Mining process. Consequently, the raw data of each set is preprocessed and created for the identification model by applying three algorithms as previously mentioned. Finally, this research uses the comparison of Accuracy and ROC Curve results as Model Evaluation.

Evaluation method : Evaluation method [11] in this research used the 10 fold cross validation, which measures the performance of the model in order to predict the sample. The basis of this technique is sampling. Firstly, data is divided into sections called fold, then putting some of that data as training set for training the model and testing set for testing and evaluation the predictive model. Ten- fold cross validation process as shown in Figure 3.5.

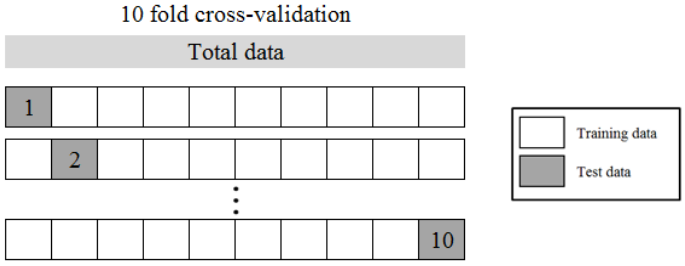


Figure 3.5 Ten- fold cross validation.

Accuracy : Accuracy is percentage value of testing set examples correctly classified by the classifier. Generally, the model that given high accuracy value may provide more precision than the lower accuracy.

Table 3.4 Confusion Matrix.

		Predicted Class	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	TP (true positive)	FN (false negative)
	Class=No	FP (false positive)	TN (true negative)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1)$$

Receiver Operating Characteristic (ROC) Curve : ROC Curve [19] is a fundamental tool for diagnostic test evaluation. This curve is generated by plotting the true positive rate (TP) against the false positive rate (FP) at various threshold settings. The true-positive rate is also known as sensitivity or the sensitivity, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as (1 - specificity). The ROC curve is thus the sensitivity as a function off all-out.

Sensitivity or TP rate

$$TPR = TP/(TP+FN) \quad (3.2)$$

Specificity or FP rate

$$FPR = FP/(FP + TN) \quad (3.3)$$

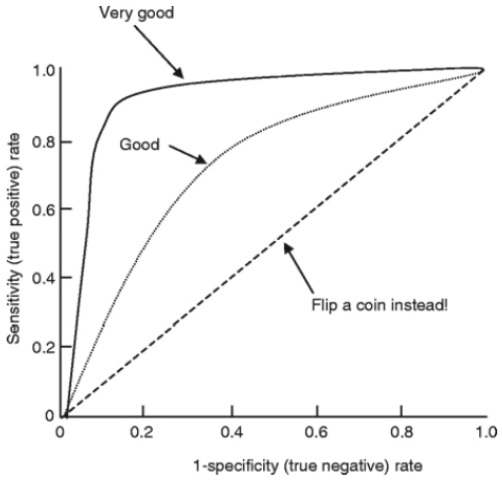


Figure 3.6 Receiver Operating Characteristic (ROC) Curve [19].

Diagnostic tools have high sensitivity and high specificity, which the latter will have low false positive rate the ROC curve resulting corner on the left. In addition, the creation of ROC curve helps to compare the efficiency of diagnosis by comparing the area under the curve of each test. The Area Under curve (AUC) represents a more efficient diagnosis.

3.2 Research Schedule

The steps and operating methods of research cloud explain as Figure 3.7.

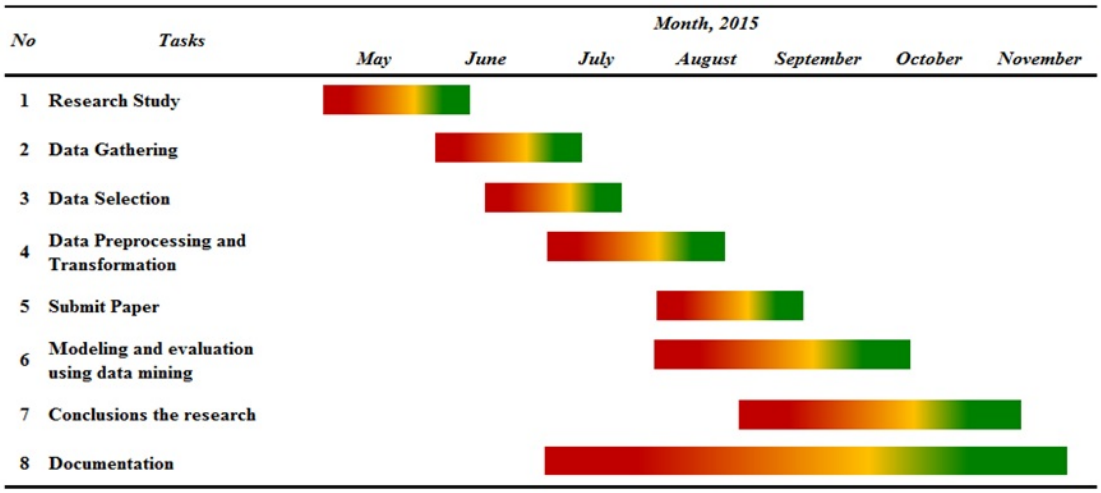


Figure 3.7 Scheduling of Research.

This chapter is methodology of this research. Next chapter will discuss the experimental results taken from the processing of data.

CHAPTER IV
EXPERIMENTAL RESULTS

According to the knowledge discovery approaches discussed in the previous section, this section presents the experimental results of classification system.

4.1 Results of Diabetes Dose Titration Identification System.

To build the Dose Titration Identification Model, this research is proceeded by three selecting classifier algorithms given as the Decision Tree, Naïve Bayes, and ANN. The related factors considered in comparison are the Accuracy and ROC Curve ratio. The results are as reported in Tables 4.1 - 4.4.

Table 4.1 Accuracy determined by Decision Tree, Naïve Bayes, and ANN.

Dataset	Medicine	Accuracy		
		Decision Tree	***Naïve Bayes***	***ANN*** *HL a, LR 0.3, MT 0.2*
Independent Dose Titration (IDT)	Metformin	0.62	0.42	0.52
	Sulfonylurea	0.51	0.40	0.43
	Thiazolidinedione	0.75	0.45	0.62
	Insulin	0.41	0.38	0.38
Historical Dose Titration (HDT)	Metformin	0.56	0.44	0.70
	Sulfonylurea	0.43	0.41	0.55
	Thiazolidinedione	0.67	0.48	0.84
	Insulin	0.40	0.40	0.43

Table 4.2 Order of Accuracy compared between IDT and HDT.

Algorithms	Metformin		Sulfonylurea		Thiazolidine-dione		Insulin	
	IDT	***HDT***	***IDT***	***HDT***	***IDT***	***HDT***	***IDT***	***HDT***
Decision Tree	0.62	0.56	0.51	0.43	0.75	0.67	0.41	0.40
Naïve Bayes	0.42	0.44	0.40	0.41	0.45	0.48	0.38	0.40
ANN	0.52	0.70	0.43	0.55	0.62	0.84	0.38	0.43

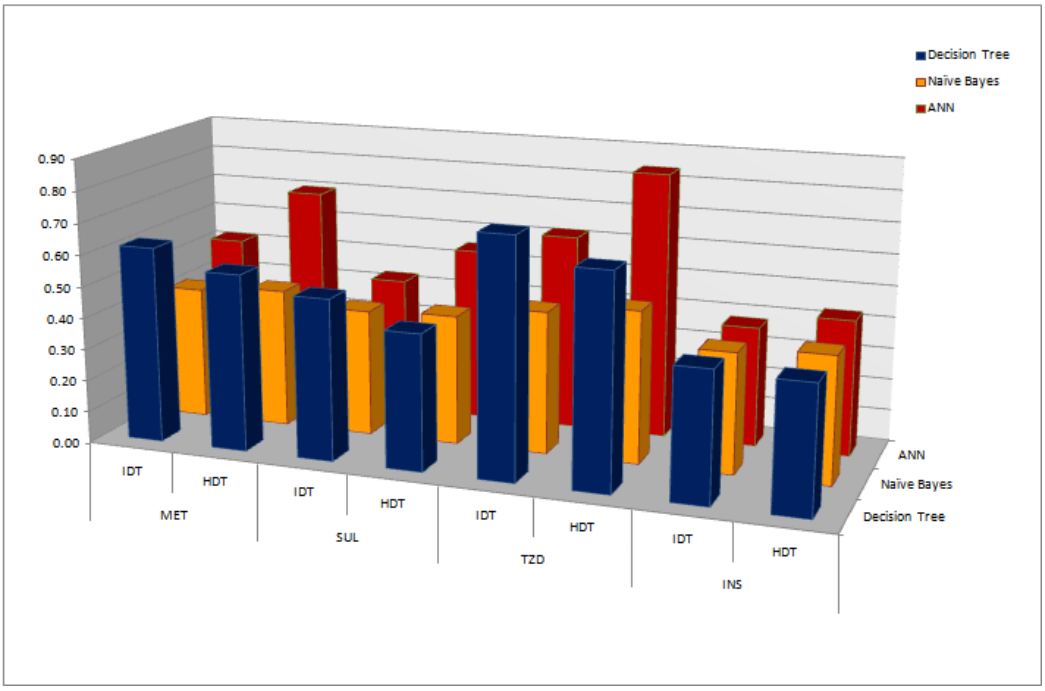


Figure 4.1 Accuracy determined by Decision Tree, Naïve Bayes, and ANN

Table 4.1 shows accuracy of IDT dataset and HDT dataset by running the same datasets of four drugs, namely, Metformin, Sulfonylurea, Thiazolidinedione and Insulin through the three algorithms: Decision Tree, Naïve Bayes, and ANN. Decision Tree provides the highest accuracy of IDT dataset. Remarkably, the order of the success rate is Thiazolidinedione (0.75), Metformin (0.62), Sulfonylurea (0.51) and Insulin (0.41). Meanwhile, ANN provides the highest accuracy of HDT dataset.

Table 4.2 reveals that there is an order of the accuracy rate of each drug dataset of IDT run through Decision Tree. The highest is Thiazolidinedione (0.75), followed by Metformin (0.62), Sulfonylurea (0.51), and the lowest is Insulin (0.41). The order is the same as that of HDT dataset run through ANN. That is Thiazolidinedione (0.84), followed by Metformin (0.70), Sulfonylurea (0.55), and the lowest is Insulin (0.43). Both pieces of information reveal that the success rate of IDT and HDT derived from Decision Tree and ANN follow the same pattern of accuracy order is shown in Figure 4.1

Table 4.3 ROC Curve of Decision Tree, Naïve Bayes, and ANN.

Dataset	**Medicine**	**ROC Curve**		
		Decision Tree	***Naïve Bayes***	***ANN*** *HL a, LR 0.3, MT 0.2*
Independent Dose Titration (IDT)	Metformin	0.78	0.59	0.69
	Sulfonylurea	0.71	0.57	0.61
	Thiazolidinedione	0.88	0.64	0.79
	Insulin	0.60	0.56	0.57
Historical Dose Titration (HDT)	Metformin	0.74	0.63	0.84
	Sulfonylurea	0.60	0.60	0.74
	Thiazolidinedione	0.84	0.66	0.91
	Insulin	0.55	0.57	0.62

Table 4.4 Order of ROC Curve compared between IDT and HDT.

Algorithms	**Metformin**		**Sulfonylurea**		**Thiazolidine-dione**		**Insulin**	
	IDT	***HDT***	***IDT***	***HDT***	***IDT***	***HDT***	***IDT***	***HDT***
Decision Tree	0.78	0.74	0.71	0.60	0.88	0.84	0.60	0.55
Naïve Bayes	0.59	0.63	0.57	0.60	0.64	0.66	0.56	0.57
ANN	0.69	0.84	0.61	0.74	0.79	0.91	0.57	0.62

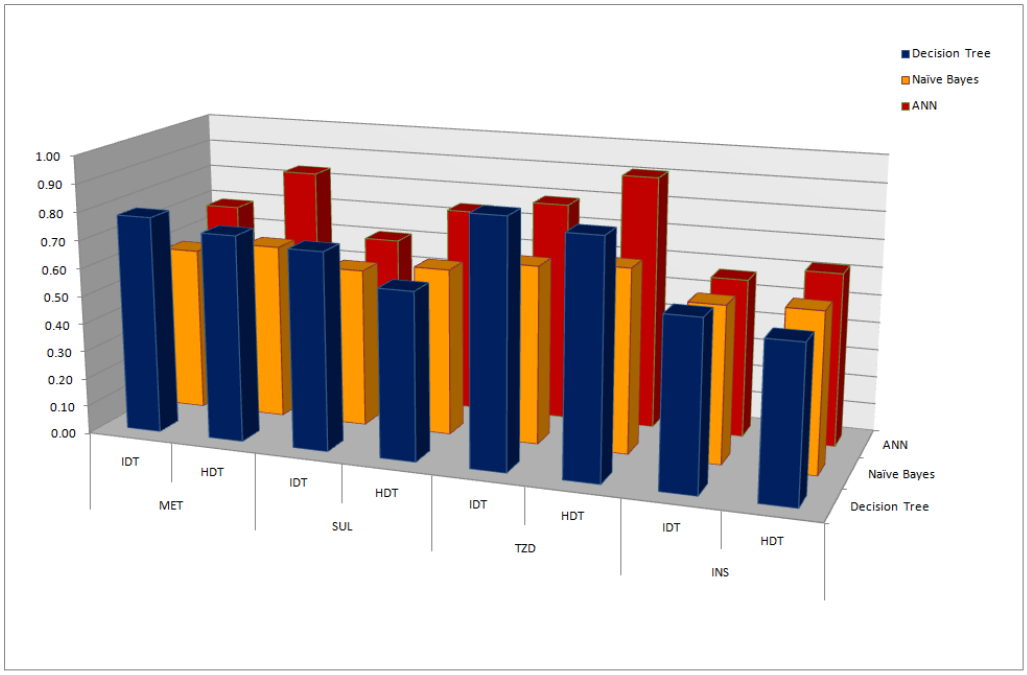


Figure 4.2 ROC Curve of Decision Tree, Naïve Bayes, and ANN.

Table 4.3 shows ROC Curve of IDT and HDT by running the same datasets of four drugs, namely, Metformin, Sulfonylurea, Thiazolidinedione and Insulin through the three algorithms: Decision Tree, Naïve Bayes, and ANN. Decision Tree provides the highest ROC Curve result of IDT. Remarkably, the order of the success rate is Thiazolidinedione (0.88), Metformin (0.78), Sulfonylurea (0.71) and Insulin (0.60). Meanwhile, ANN provides the highest ROC Curve result of Historical HDT.

Table 4.4 reveals that there is an order of the ROC Curve rate of each drug dataset of IDT run through Decision Tree. The highest is Thiazolidinedione (0.88), followed by Metformin (0.78), Sulfonylurea (0.71), and the lowest is Insulin (0.60). The order is the same as that of HDT run through ANN. That is Thiazolidinedione (0.91), followed by Metformin (0.84), Sulfonylurea (0.74), and the lowest is Insulin (0.62). Both pieces of information reveal that the success rate of IDT and HDT derived from Decision Tree and ANN follow the same pattern of accuracy order is shown in Figure 4.2.

In summary, by comparing only the accuracy results of 2 datasets: IDT and HDT, run through three algorithms, HDT processed by ANN outperforms IDT. Since the experimental result of this model maintain as default parameter, it possibly to get more efficient model if default parameter was adjusted which be explained in next step.

4.2 Parameter adjustment for ANN algorithms.

There are only 3 parameters that are consequently adjusted in this experiment only for HDT dataset. They are Hidden Layer (HL), Learning Rate (LR), and Momentum (MT), which are applied to obtain a more trustworthy and more efficient results model. The experiment results are shown in Table 4.5.

Table 4.5 Post adjustment results of ANN Algorithms.

ANN Algorithms	**Metformin** *HL 34, LR 0.3, MT 0.2*		**Sulfonylurea** *HL 32, LR 0.3, MT 0.2*		**Thiazolidine dione** *HL 32, LR 0.275, MT 0.2*		**Insulin** *HL 32, LR 0.3, MT 0.2*	
	Acc	**ROC**	**Acc**	**ROC**	**Acc**	**ROC**	**Acc**	**ROC**
Default	0.70	0.85	0.56	0.75	0.85	0.91	0.44	0.63
Post Adjustment	0.82	0.92	0.65	0.88	0.91	0.95	0.46	0.51
% Change	**17.14**	**8.24**	**16.07**	**17.33**	**7.06**	**4.40**	**4.55**	**-19.05**

Table 4.5 showing the comparison results between pre-adjustment (Default) and post-adjustment. Thiazolidinedione provids the highest accuracy of 91% and ROC Curve result reaches 95% by tuning the parameters: Hidden Layer(HL) 32,Learning Rate (LR) 0.275, and Momentum (MT) 0.2, respectively.

However, the result of Insulin value is still low, although the parameter has been adjusted. Thus, this research brings the Insulin HDT dataset to preprocess prior to building the efficient model. This will be shown in the next topic.

4.3 Data reduction and aggregation for Insulin dose Titration

This research uses the method of data reduction and aggregation to eliminate the size of data and processing time.

Details of Insulin HDT dataset is shown in Table 4.6, example of Insulin HDT dataset is shown in Figure 4.3 and the results of Insulin HDT dataset is shown in Table 4.7.

Table 4.6 Data dictionary of Insulin HDT dataset.

Feature name	Description
Gender	Values: male, female
Age	Range of patients age; 40, 50, 60, …, 90
Diagnosis 1	The primary diagnosis.
A1c test result	Value : 0,1,2, and 3
MetNew	Historical record of Metformin titration was added by two previous times retrospectively including the recent record. Value statuses are as follows : S = Single Steady, SU = Steady and Up, SD = Steady and Down, and V = Variance.
SulNew	Historical record of Sulfonylurea titration was added by two previous times retrospectively including recent record. Value statuses are as follows : S = Single Steady, SU = Steady and Up, SD = Steady and Down, and V = Variance.
TZDNew	Historical record of Thiazolidinedione titration was added by two previous times retrospectively including the recent record. Value statuses are as follows : S = Single Steady, SU = Steady and Up, SD = Steady and Down, and V = Variance.

Table 4.6 Data dictionary of Insulin HDT dataset. (Cont).

Feature name	Description
InsNew	Historical record of Insulin titration was added by two previous times retrospectively. Value statuses are as follows : S = Single Steady, SU = Steady and Up, SD = Steady and Down, and V = Variance.
Ins	Insulin prescription record. Value statuses are as follows : S = Single Steady, SU = Steady and Up, SD = Steady and Down, and V = Variance.

No.	gender Nominal	age Numeric	diag_1 Numeric	A1Cresult Numeric	MetNew Nominal	SulNew Nominal	TZDNew Nominal	InsNew Nominal	**Ins** Nominal
1	Female	80.0	2.0	0.0				S	Down
2	Male	60.0	9.0	0.0				S	Down
3	Male	60.0	3.0	0.0	S			S	Steady
4	Female	70.0	1.0	3.0		SU		S	Down
5	Female	80.0	1.0	1.0				S	Steady
6	Male	50.0	1.0	3.0	S	S		S	Up
7	Female	80.0	3.0	0.0				S	Steady
8	Male	60.0	2.0	1.0	S			S	Steady
9	Female	70.0	1.0	0.0	S		S	S	Steady
10	Female	50.0	2.0	1.0	S	S		S	Steady
11	Female	50.0	1.0	3.0	S	S		S	Up
12	Female	60.0	1.0	2.0				S	Down
13	Male	50.0	3.0	0.0	S		S	S	Steady
14	Female	50.0	1.0	1.0		S	S	S	Steady
15	Female	40.0	1.0	2.0				S	Steady
16	Female	50.0	2.0	0.0		S		S	Steady

Figure 4.3 Example of Insulin HDT dataset.

Table 4.7 Results of Insulin HDT dataset operating by ANN algorithms.

Insulin Dose Tritration	Accuracy	ROC
Before preprocess *HL 32, LR 0.3, MT 0.2*	0.46	0.51
After preprocess *HL 30, LR 0.275, MT 0.2*	0.70	0.82
% Change	**52.17**	**60.78**

Results in Table 4.7 is presented that ANN Model could generate the higher accuracy upto 0.70 and ROC Curve results upto 0.82 after preprocessing the Insulin HDT dataset.

In summary, the best results is HDT dataset that could be generated model with the most accuracy result when performing on ANN algorithms based on the default parameters. When parameters are adjusted by Hidden Layer (HL), Learning Rate (LR), and Momentum, the accuracy and ROC Curve results are higher value than default parameters except the Insulin. After that Insulin HDT dataset was preprocess again before building up model and getting the highest results for both accuracy and ROC Curve. The accuracy reach up to 0.70 and ROC Curve reach up to 0.82. It implies that the means data preprocessing and parameter adjustment could effect on model efficiency.

This chapter are experimental results of research. Next chapter conclusions, discussion and future works are proposed.

CHAPTER V
CONCLUSION AND FUTURE WORKS

This chapter focuses the use of classification techniques by Decision tree, Naïve Bayes, and ANN for analysis historical data of diabetes patients. The classification model is proposed to identify the diabetic medication adjustment. By the experimental results, the discussion is also given, conclusion and suggestion are also described as follows.

5.1 Research Conclusion

This researcher presents the classification techniques for applying the Diabetes dose titration identification model to support the decision making in dose titration identification of type-2 diabetes inpatients. Technically, the classification methods of Decision Tree, Naïve Bayes, and ANN are considered in this research. Data in this research is extracted from Health Fact Database in 130 hospitals in the United States and the primary raw data is preprocessed by resampling, then it is categorized into 2 datasets including IDT dataset and HDT dataset. Then each dataset is sent out for model building, model evaluation, and eventual deployment. Regarding to the experimental results considered by the accuracy and ROC Curve, Decision Tree is the most appropriate method for IDT dataset, because it generates higher accuracy and ROC Curve value than other methods. On the other hand, ANN Algorithm generates the model with high accuracy and ROC Curve for HDT dataset.

In summary, the best results is HDT dataset which could be generated with the most accuracy result when performing by ANN algorithms based on default parameter. When Hidden Layer (HL),Learning Rate (LR), and Momentum are adjusted, the accuracy and ROC Curve results are better than the defaults parameter except Insulin. Then, the Insulin HDT dataset refeedbacks to preprocess again before building up model and get the highest results interms of accuracy and ROC Curve.

The accuracy reaches up to 0.70 and ROC Curve reaches up to 0.82. Therefore, the best method of Diabetes Dose Titration Identification Model should be considered by being based on the proper used dataset, data preprocessing, and parameter adjustment.

According to classification system, the selected classifier methods in this research are obviously proved that the system is effective and should be considered as the preliminary process to identify dose titration, supporting the medical decision making and treatment planning that suits for individual type-2 diabetic inpatients.

5.2 Future Works

Model creating by Decision Tree and operate with IDT Dataset could generate the highest accuracy and ROC Curve. Because IDT Dataset is the independent data and less complicated. Meanwhile, HDT Dataset when processing with ANN could provide high accuracy and ROC Curve than Decision Tree due to data is complicated.

In view of data aspect, model should be more punctual and reliable if data is enhanced by adding the combination drug and the exceeding 2 times historical dose adjustment. Nevertheless, this research may fulfill more efficient to the model if some limitation relating to inadequate variables and high missing value of data are reduced.

Insulin HDT dataset preprocessing could effect on model efficiency and may be applied with other drugs in future in order to increase the precision in identification model.

The classification model is created for proper identifying drug and dose titration in each patient. Researcher suggest to adopt Association technique for applying in cooperated with classification technique in future work. Because of the association rule could generate relevant rules to deeply define drug adjustment model. In future when the predictive model such as classification techniques is combined with the descriptive model like association rules, it should create the most efficient Diabetes Dose Titration Identification model that make more benefit in Diabetes inpatients therapy. However, if the additional historical data is provided, it could be better to adopt preprocessing data by control chart pattern method [20]. This could bring the results to apply with other field.

REFERENCES

1 Alberti, K.G., Zimmet, P.Z., 1998. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med. 15(7), 539-53.

2 International Diabetes Federation. 2014. IDF Diabetes Atlas [Online].Available from : www.idf.org/diabetesatlas [2015, June 3]

3 American Diabetes Association. 2015. Standard of medical care in diabetes. Diabetes Care. 38,(Suppl.1).

4 The Health and Social Care Information Centre. National Diabetes Audit 2010-2011 Report. Complications and Mortality, UK. [accessed 2015 June 3], Available from : www.idf.org/diabetesatlas

5 Wiraphol Phimarn, Phayom Sookaneknun. Current Principle of Pharmacotherapy in Diabetes Mellitus. Thai Pharmaceutical and Health Science Journal. 3(1), 169-179.

6 Hex N., Bartlett C., Wright D., Taylor M., Varley D. 2012. Estimating the current and future costs of Type-1 and Type-2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. Diabet Med. 29(7), 855-862.

7 The UCI Machine Learning Repository. 2014. Diabetes 130-US hospitals for years 1999-2008 [Online]. Available from : http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008[2015, February 15]

8 American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 37(Suppl. 1), S81–S90.

9 The International Expert Committee. 2009. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. Diabetes Care. 32, 1327–1334.

10 Mayo Clinic Hospital, Saint Marys Campus. 2014. Treatments and drugs [Online]. Available from : http://www.mayoclinic.org/diseases-conditions/Type-2-diabetes/basics/treatment/con-20031902 [2015, August 10]

11 Tutorials Point originated. Decision Tree - Classification. Available from :http://www.tutorialspoint.com [2015, October 10]

12 Wikimedia Foundation, Inc.2015. Cross Industry Standard Process for Data Mining[Online]. Available from : https://en.wikipedia.org/wiki/Cross_ Industry_Standard_Process_for_Data_Mining[2015, October 31]

13 Eakasit Pacharawongsakda. 2014. An Introduction to Data Mining Techniques.

14 Chinnapat Kaewchinporn. 2013. C4.5 decision tree algorithm [Online]. Available from : http://scriptslines.com/blog/category/knowledge-discovery-in-database-kdd-and-data-mining/ [2015, October 15].

15 Seksunti J., Somjit A., Ngamnij A. and Fong Tsai, Ya-Han Hu and Watchapon D. 2012. Smart Assistant System for Chronic Disease Healthcare Planning Using Data Mining. 4th National Conference on Information Technology. Phetchaburi, April 26 - 27, 2012, 117-122.

16 Cook C.B., McMichael J.P., Lieberman R., Mann L.J., King E.C., New K.M., Vaughn P.S., Dunbar V.G., Caudle J.M. 2005. The Intelligent Dosing System : Application for Insulin Therapy and Diabetes Management. Diabetes Technology & Therapeutics. 7(1).

17 Canadian Diabetes Association Clinical Practice Guidelines Expert Committee. 2013. In-hospital Management of Diabetes. Can J Diabetes. 37, S77-S81.

18 Beata S track, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, John N. Clore. 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International. 2014(2014).

19 Galley H.F. 2004. Editorial II : Solid as a ROC. Medicine & Health. 93(5), 623-626.

20 Jiemin Wang, A. K. Kochhar, R. G. Hannam. 1998. Pattern recognition for statistical process control. The International Journal of Advanced Manufacturing Technology. 14(2), 99-109.

APPENDICES

APPENDIX A

INDEX FOR 3 AND 4 DIGIT DIAGNOSTIC CODES (ICD9)

INDEX FOR 3 AND 4 DIGIT DIAGNOSTIC CODES (ICD9)

INFECTIONS AND PARASITIC DISEASES

001 - 009.3	Intestinal and Infectious Diseases
010 - 018.9	Tuberculosis
020 - 027.9	Zoonotic Bacterial Diseases
030 - 041.9	Other Bacterial Diseases
045 - 049.9	Poliomyelitis and Other Non-Arthropod Borne Viral Diseases of Central Nervous System
050 - 057.9	Viral Diseases Accompanied by Exanthem
060 - 066.9	Arthropod -Borne Viral Diseases
070 - 079.9	Other Diseases Due to Viruses and Chlamydiae
080 - 088.9	Rickettsiosis and Other Arthropod Borne Diseases
090 - 099.9	Syphilis and Other Venereal Diseases
100 - 104.9	Other Spirochaetal Diseases
110 - 118	Mycosis
120 - 129	Helminthiasis
130 - 136.9	Other Infectious and Parasitic Diseases
137 - 139.8	Late Effects of Infectious and Parasitic Diseases

NEOPLASMS

140 - 149.9	Malignant Neoplasm of Lip, Oral Cavity and Pharynx
150 - 159.9	Malignant Neoplasm of Digestive Organs and Peritoneum
160 - 165.9	Malignant Neoplasm of Respiratory and Intrathoracic Organs
170 - 175.9	Malignant Neoplasm of Bone, Connective Tissue, Skin and Breast
179 - 189.9	Malignant Neoplasm of Genitourinary Organs
190 - 199.1	Malignant Neoplasm of Other and Unspecified Sites
200 - 208.9	Malignant Neoplasm of Lymphatic and Haematopoietic Tissue
210 - 229.9	Benign Neoplasm
230 - 234.9	Carcinoma in Situ
235 - 238.9	Neoplasms of Uncertain Behaviour
239 - 239.9	Neoplasms of Unspecified Nature

ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES AND IMMUNITY DISORDERS

240 - 246.9	Disorders of Thyroid Gland
250 - 259.9	Diseases of Other Endocrine Glands
260 - 269.9	Nutritional Deficiencies
270 - 279.9	Other Metabolic Disorders and Immunity Disorders

DISEASES OF BLOOD AND BLOOD FORMING ORGANS

280 - 289.9	Diseases of Blood and Blood Forming Organs

MENTAL DISORDERS

290 - 294.9	Organic Psychotic Conditions
29.5 - 299.9	Other Psychoses
300 - 316	Neurotic Disorders, Personality Disorders and Other Nonpsychotic Mental Disorders
317 - 319	Mental Retardation

DISEASES OF NERVOUS SYSTEM AND SENSE ORGANS

320 - 326	Inflammatory Diseases of the Central Nervous System
330 - 337.9	Hereditary and Degenerative Diseases of Central Nervous System
340 - 349.9	Other Disorders of the Central Nervous System
350 - 359.9	Disorders of the Peripheral Nervous System
360 - 379.9	Disorders of the Eye and Adnexa
380 - 389.9	Disorders of the Ear and Mastoid Process

DISEASES OF THE CIRCULATORY SYSTEM

390 - 392.9	Acute Rheumatic Fever
393 - 398.9	Chronic Rheumatic Heart Disease
401 - 405.9	Hypertensive Disease
410 - 414.9	Ischaemic Heart Disease
415 - 417.9	Diseases of Pulmonary Circulation
420 - 429.9	Other Forms of Heart Disease
430 - 438	Cerebrovascular Disease
440 - 448.9	Diseases of Arteries, Arterioles and Capillaries
451 - 459.9	Diseases of Veins and Lymphatics, and Other Diseases of Circulatory system

DISEASES OF THE RESPIRATORY SYSTEM

460 - 466.1	Acute Respiratory Infections
470 - 478.9	Other Diseases of Upper Respiratory Tract
480 - 487.8	Pneumonia and Influenza
490 - 496	Chronic Obstructive Pulmonary Disease and Allied Conditions
500 - 508.9	Pneumoconioses and Other Lung Diseases due to .External Agents
510 - 519.9	Other Diseases of Respiratory System

DISEASES OF THE DIGESTIVE SYSTEM

520 - 529.9	Diseases of Oral Cavity, Salivary Glands and Jaws
530 - 537.9	Diseases of Oesophagus, Stomach and Duodenum
540 - 543	Appendicitis
550 - 553.9	Hernia of Abdominal Cavity
555 - 558	Noninfective Enteritis and Colitis
560 - 569.9	Other Diseases of Intestines and Peritoneum
570 - 579.9	Other Diseases of Digestive System .

DISEASES OF GENITOURINARY SYSTEM

580 - 589.9	Nephritis, Nephrotic syndrome and Nephrosis
590 - 599.9	Other Diseases of Urinary System
600 - 608.9	Diseases of Male Genital Organs
610 - 611.9	Disorders of Breast
614 - 616.9	Inflammatory Disease of Female Pelvic Organs
617 - 629.9	Other Disorders of Female Genital Tract

COMPLICATIONS OF PREGNANCY, CHILDBIRTH AND THE PUERPERIUM

630 - 639.9	Pregnancy with Abortive Outcome
640 - 648.9	Complications Mainly Related to Pregnancy
650 - 659.9	Normal Delivery and Other Indications for Care in Pregnancy, Labour and Delivery
660 - 669.9	Complications Occurring Mainly in the Course of Labour and Delivery
670 - 676.9	Complications of the Puerperium

DISEASES OF THE SKIN AND SUBCUTANEOUS TISSUE

680 - 686.9	Infections of Skin and Subcutaneous Tissue
690 - 698.9	Other Inflammatory conditions of Skin and Subcutaneous Tissue
700 - 709.9	Other Diseases *of* Skin and Subcutaneous Tissue

DISEASES OF MUSCULOSKELETAL SYSTEM AND CONNECTIVE TISSUE

710 - 719.9	Arthropathies and Related Disorders
720 - 724.9	Dorsopathies
725 - 729.9	Rheumatism, Excluding the Back
730 - 739.9	Osteopathies, Chondropathies and Acquired Musculoskeletal Deformities

CONGENITAL ANOMALIES

740 - 759.9	Congenital Anomalies

CERTAIN CONDITIONS ORIGINATING IN THE PERINATAL PERIOD

760 - 779.9	Certain Conditions Originating in the Perinatal Period

SYMPTOMS, SIGNS AND ILL-DEFINED CONDITIONS

780 - 789.9	Symptoms
790 - 796.9	Nonspecific Abnormal Findings
797 - 799.9	Ill-defined and Unknown Causes of Morbidity and Mortality

INJURY AND POISONING

800 - 804.3	Fracture of Skull
805 - 809.1	Fracture of Spine and Trunk
810 - 819.1	Fracture of Upper Limb
820 - 829.1	Fracture of Lower Limb
830 - 839.9	Dislocation
840 - 848.9	Sprains and Strains of Joints and Adjacent Muscles
850 - 854.1	Intracranial Injury Excluding those with Skull Fractures
860 - 869.1	Internal Injury of Chest, Abdomen and Pelvis
870 - 879.9	Open Wound of Head, Neck and Trunk
880 - 887.7	Open Wound of Upper Limb
890 - 897.7	Open Wound of Lower Limb
900 - 904.9	Injury to Blood Vessels
905 - 909.9	Late Effects of Injuries, Poisonings, Toxic Effects and Other External Causes
910 - 919.9	Superficial Injury
920 - 924.9	Contusion with Intact Skin Surface
925 - 929.9	Crushing Injury
930 - 939.9	Effects of Foreign Body Entering Through Orifice
940 - 949.9	Burns
950 - 957.9	Injury to Nerves and Spinal Cord
958 - 959.9	Certain Traumatic Complications and Unspecified Injuries
960 - 979.9	Poisoning by Drugs, Medicaments and Biological Substances
980 - 989.9	Toxic Effects of Substances Chiefly Nonmedical as to Source
990 - 995.8	Other and Unspecified Effects of External Causes
996 - 999.9	Complications of Surgical and Medical Care Not Elsewhere Classified

SUPPLEMENTARY CLASSIFICATIONS OF FACTORS INFLUENCING HEALTH STATUS AND CONTACT WITH HEALTH SERVICES (V01 - V82)

V0l - V07.9	Persons with Health Hazards Related to Communicable Diseases
V10 - V19.8	Persons with Potential Health Hazards Related to Personal and Family History
V20 - V28.9	Persons Encountering Health Services in circumstances Related to Reproductior and Development
V30 - V39.2	Healthy Liveborn Infants According to Type of Birth
V40 - V49.9	Persons with Conditions Influencing Their Health status
V50 - V59.9	Persons Encountering Health Services for Specific Procedures and After
V60 - V68.9	Persons Encountering Health Services in Other Circumstances
V70 - V82.9	Persons without Reported Diagnosis Encountered during Examination and Investigation of Individuals and Populations

ADDITIONAL DIAGNOSTIC CODES

01A	Dizziness, Vertigo, Insomnia
02A	Abdominal Swelling Not Otherwise Specified or Abdominal Pain
03A	Pre-Operative Assessment (Dental) – No Diagnosis Specified
04A	General Psychiatric Examination – No Care Required
05A	Growth and Development
06A	Feeding and Management Talk/Anxiety of Mother
07A	Feeding Problem
08A	Healthy Newborn Care
10A	Emergency Care – Assault
11A	Nothing Abnormal Discovered
12A	Epistaxis/Cautery
31A	Removal of Sutures
32A	Injection – Allergy
33A	Injection – Other
34A	Contraceptive Advice
35A	Benign Skin Lesions Including Keratosis, Warts other than Plantar Warts (for Plantar Warts See 45a)
36A	Fecal Impaction
42A	Removal of Cast
43A	Change of Dressing
44A	Contact with Communicable Diseases
45A	Plantar Warts
01B	Tuberculosin Skin Test
02B	Skin Grafting
03B	Keloid Scarring
06B	Syringing of Ears
08B	Congenital Anomalies of the Lower Respiratory System
10B	Consultation Re Sterilization – Male
11B	Genetic Counselling – Male
12B	Sterilization – Male
15B	Sterilization – Female
16B	Consultation Re Sterilization – Female

17B	Consultation Re Abortion
18B	Genetic Counselling – Female
19B	Artificial Insemination
20B	Post Coital Test
21B	Congenital Anomalies – Female
22B	Open Wounds of Genital Organs – Female
23B	Insertion/Removal of IUD
30B	Prenatal Care
31B	Hypertrophy of Breast, Mammary Gland, Nipple Arising During Pregnancy
32B	Erosion and Inflammation Of Cervix (Uteri) Arising During Pregnancy
33B	Leukerrhea, Vaginal Discharge Not Otherwise Specified Arising During Pregnancy
34B	Hypertensive Disease Arising During Pregnancy
35B	False Labour
36B	Pregnancy, Examination Pregnancy Unconfirmed
37B	Premature Rupture of Membranes
38B	Threatened Abortion
50B	Anxiety/Depression
55B	Foreign Body, Hand or Finger
60B	Foreign Body, Foot or Toes
65B	Animal Bite
66B	Insect Bite
01E	Eye Tests
01F	Ear Tests
01H	Hospital
01L	Laboratory
01X	X-Ray
01Z	Anaesthetic
E01	High Refractive Error (+/-8 Dioptre Or More)
E02	Change of 0.5 Dioptres or > to Spherical or Cylinder Lens.
E03	0.5 Dioptres or Greater Change to Cylinder Lens
E04	Change in Axis = > Cylinder Lens of .5 Dioptres and <20 Degree
E05	Change in Axis Of =/> 20 Degrees for a Cylinder Lens of 0.5 Dioptre or <
E06	10 Degrees for a Cylinder Lens of >0.5 Dioptre but not >1.0 Dioptre
E07	Intraocular Surgery
E08	Medications
E09	3 Degrees for a Cylinder Lens of more than 1.0 Dioptre
E10	Previously +/- 8 D Or Greater, at Risk of Retinal Detachment
E91	'No' Indicator Present
E92	Indicator of Ocular Pathology: External
E93	Indicator of Ocular Pathology: Internal
E94	Indicator of Binocularity: Phoria
E95	Indicator of Binocularity: Strabismus
E96	Indicator of Vision: Amblyopia
E97	Indicator of Refractive Error: Astigmatism
E98	Indicator of Refractive Error: Hyperopia
E99	Indicator of Refractive Error: Myopia

APPENDIX B

THE PROCEEDINGS OF THE 8th BIOMEDICAL ENGINEERING INTERNATIONAL CONFERENCE 2015 (BMEiCON 2015)

The 2015 Biomedical Engineering International Conference (BMEiCON-2015)

Diabetes Dose Titration Identification Model

Ratchanee Kaewthai[1], Sotarat Thammaboosadee[2], Supaporn Kiattisin[3]
Technology of Information System Management Division,
Faculty of Engineering, Mahidol University, Nakhonpathom, Thailand
[1]ratcha.ff@hotmail.com, [2]sotarat.tha@mahidol.ac.th, [3]supaporn.kit@mahidol.ac.th

***Abstract*— Diabetes is a chronic disease that requires continuous treatment throughout lifespan and increased risk opportunity of developing a number of serious health problems, which are high treatment cost. Admitted diabetes inpatients should receive the appropriate treatment in order to reduce rating of severe complications and premature death. This paper aims to develop the classification model for diabetic medication adjustment based on historical medical record of diabetic inpatients by applying three algorithms; Decision Tree, Naïve Bayes and Artificial neural network. By comparison of the results of each method, Decision Tree is outperformed than others for Independent Dose Titration Model (IDT) dataset and Artificial Neural Network algorithm generated model with high accuracy and ROC Curve for Historical Dose Titration Model (HDT) dataset. The results of this paper could support the decision making in medication adjustment of diabetes inpatients, particularly type-2 diabetes inpatients.**

***IndexTerms*— Dose titration; Diabetes inpatient; Data Mining; Identification Model; Medical Management .**

I. Introduction

Diabetes is caused by high glucose level of blood when the pancreas does not produce high enough insulin to handle the storage of sugar from the blood into the cells to keep blood sugar levels return to normal [1]. The current global diabetes populations are about 387 million people and trend to be increased to 592 million people within 2035. Total 90% of worldwide diabetic patients are type-2 diabetes. According to the standard of American Diabetes Association Standards of Medical Care in Diabetes [2], Diabetes care in the hospital planning should start at hospital admission from a collaborative, integrated team with expertise in diabetes. Diabetes inpatients were classified by critical level into two types; Critically ill patients and Non-critically ill patients [3]. The main purpose of treatment is to control and maintain glycemic within the appropriate levels. Medically, there are two important types of medication which are insulin injections and oral questions.

Currently, there are still no supported researches to find out that which one of the both mentioned therapies is the best because it depends on expertise's decision only [4]. Thus, it is interesting to develop the model by applying data mining techniques with the diabetic history in order to adjust the proper drug dosage and provide precise medication planning for the individual. The acquired model could help the relevant expert team to effectively deliver more accurate treatment than existing and the past.

This paper provides a data mining task for intelligent data analysis in medical field approaching to a machine learning perspective based on the historical data of diabetic inpatients. The classification models are proposed to identification the diabetic medication adjustment. The objective of the proposed model is to support the medical decision making and treatment planning that suits for the individual diabetic inpatients.

This paper organized as follow: section II presents related research, section III and IV present methodology and experimental results respectively. Finally, the conclusion is presented.

II. Related Research

A. Type-2 Diabetes

Diabetes is a chronic condition in which the levels of glucose in the blood are too high. Blood glucose levels are normally regulated by the hormone insulin, which is produce by the pancreas. In people with diabetes, the pancreas does not produce enough insulin or there is a problem with how the body's cells respond to it. This causes the patient to be severe hyperglycemia. Diabetes is a chronic and incurable disease such as heart disease, stroke, kidney failure, blindness and lower limb amputation [5,6].

Type-2 diabetes are caused due to a progressive insulin secretory defect on the background of insulin resistance, This form, previously referred to as "noninsulin-dependent diabetes" or "adult onset diabetes", accounts for ; 90–95% of all diabetes. Type-2 diabetes can cause serious health complications such as heart and blood vessel disease, nerve damage (neuropathy), kidney damage (nephropathy), eye damage, foot damage, hearing impairment, skin conditions and Alzheimer's disease.

Diabetes may be diagnosed based on Hemoglobin A1c (A1c) criteria or plasma glucose criteria, either the fasting plasma glucose (FPG) or the 2-h plasma glucose (2-h PG) value after a 75-g oral glucose tolerance test (OGTT) [7,8]. Testing to detect type-2 diabetes in asymptomatic people should be considered in adults of any age who are overweight or obese and testing should begin at age 45 years. If tests are normal, repeat testing carried out at a minimum of 3-year intervals is reasonable [3].

978-1-4673-9158-0/15/$31.00 ©2015 IEEE

B. Pharmacological Therapy and medication management

Type-2 diabetes can achieve their target blood sugar levels with diet and exercise alone, but most of them also need diabetes medications or Insulin therapy. The decision about which medications are the best depends on many factors, including blood sugar level and any other health problems the patient may have. The doctor might even combine drugs from different classes to control glycemic in various different ways [9].

Examples of possible treatments for type-2 diabetes include [9]:

- **Metformin**: It is the first medication prescribed for type-2 diabetes. It works by improving the sensitivity of your body tissues to insulin so that your body uses insulin more effectively.
- **Sulfonylureas:** It is medications help your body secrete more insulin. Examples of medications in this class include Glyburide, Glipizide and Glimepiride. Possible side effects include low blood sugar and weight gain.
- **Meglitinides:** These medications work like Sulfonylureas by encouraging the body to secrete more insulin, but they're faster acting, and they don't stay active in the body for as long. They also have a risk of causing low blood sugar, but not as much risk as Sulfonylureas do. Weight gain is a possibility with this class of medications as well. Examples include Repaglinide and Nateglinide.
- **Thiazolidinediones**: Like Metformin, these medications make the body's tissues more sensitive to insulin. This class of medications has been linked to weight gain and other more serious side effects, such as an increased risk of heart failure and fractures. Because of these risks, these medications generally aren't a first-choice treatment. Rosiglitazone and Pioglitazone are examples of Thiazolidinediones.
- **DPP-4 inhibitors**: These medications help reduce blood sugar levels, but tend to have a modest effect. They don't seem to cause weight gain. Examples of these medications are Sitagliptin, Saxagliptin and Linagliptin.
- **GLP-1 receptor agonists**: These medications slow digestion and help lower blood sugar levels, though not as much as Sulfonylureas. This class of medications isn't recommended for use alone. Exenatide and Liraglutide are examples of GLP-1 receptor agonists. Possible side effects include nausea and an increased risk of pancreatitis.
- **SGLT2 inhibitors**. These are the newest diabetes drugs on the market. They work by preventing the kidneys from reabsorbing sugar in the blood. Instead, the sugar is excreted in the urine.
- **Insulin therapy.** Some people who have type-2 diabetes need Insulin therapy as well. In the past, Insulin therapy was used as last resort, but today it's often prescribed sooner because of its benefits.

C. Medication management for diabetes inpatients

Diabetes discharge planning should start at hospital admission, and clear diabetes management instructions should be provided at discharge. The sole use of sliding scale insulin (SSI) in the inpatient hospital setting is strongly discouraged. All patients with diabetes admitted to the hospital should have their diabetes type clearly identified in the medical record. Inpatient, there are two important types of critical which are critically ill patients and non-critically ill patients [3].

Critically Ill Patients: Insulin therapy should be initiated for treatment of persistent hyperglycemia starting at a threshold of no greater than 180 mg/dl. Once Insulin therapy is started, a glucose range of 140–180 mg/dl is recommended for the majority of critically ill patients. More stringent goals, such as 110–140 mg/dl, may be appropriate for selected patients, as long as this can be achieved without significant hypoglycemia. Critically ill patients require an intravenous insulin protocol that has demonstrated efficacy and safety in achieving the desired glucose range without increasing risk for severe hypoglycemia [3].

Non-critically Ill Patients: In case treated with insulin, generally premeal blood glucose targets of <140 mg/dl with random blood glucose <180 mg/dl are reasonable, provided these targets can be safely achieved. More stringent targets may be appropriate in stable patients with previous tight glycemic control. Less stringent targets may be appropriate in those with severe comorbidities.

D. Related research

Seksunti et al [10]. Proposed research about Diabetes inpatient's knowledge discovered by Data Mining. Database of 12,000 chronic diabetes inpatient at Nongbua-Rawae Hospital during year 2006-2010 were used as a case study. Researchers analyzed patient's data by applying classification algorithms; Decision Tree with K-fold Cross Validation and Train-Test validation in mining knowledge Regarding to the model evaluation, the acquired experimental results were generated 12 rules with 83% of the accuracy. Nevertheless this work was proposed classification model in drug prescription using only patient's records. As a result the author propose new methodology to create predictive model for Diabetes titration by considering the individual drug dosage, lab results and historical therapy records as the database to build up model. With the combination of mentioned data, the acquired model is more accuracy and better support decision making in Dose Titration for type-2 diabetes inpatients.

Cook et al. [11] introduced the Intelligent Dosing System; the mathematical model that using dose response data to calculate new dose and applied to insulin adjustment for diabetes patient therapy. Retrieved data was derived from an electronic Diabetes Patient Tracking System of a large urban outpatient diabetes clinic in between year 1991 -2001. This work proposed method by based on mathematical model only, no related to Data Mining Techniques which is more effective, more complex data and provide better predictive results for in-hospital dose adjustment of Diabetes management.

Canadian Diabetes Association Clinical Practice Guidelines Expert Committee [12] was the Clinical Practice Guidelines for In-Hospital Management of Diabetes. This research guided that for in-hospital patients, blood sugar level's controlling is more significant than food controlling and diabetes management. In-hospital Management of Diabetes is used as the Clinical Practice Guidelines and be categorized in 10 main topics covering all managerial aspects about in-hospital diabetes patients. Researcher proposed many theories for being guidelines in effective dose titration management.

III. Methodology

CRISP-DM Model, as shown in Figure 1, the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. These phases help model understand the data mining process and provide a road map to follow while planning and carrying out a data mining project. This article explores all six phases, including the tasks involved with each phase. Sidebar material, which takes a look at specific data mining problem types and techniques for addressing them, is provided.

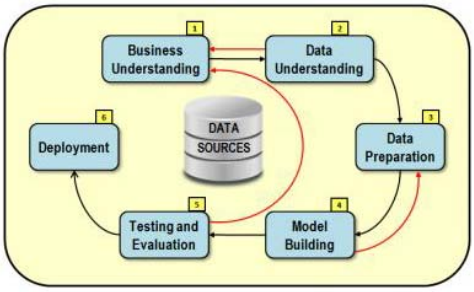


Fig. 1. Cross-Industry Standard Process for Data Mining Model.

Decision Tree [13] is one of supervised learning algorithm that widely used in predictive classification model. This model is represented by tree diagram structure. Leaves represent class labels and branches represent correlation of attributes which lead to each class labels. By tree based representation, the decision tree could explicitly interpret, describe the results and enable user to analyze data and make decision more precisely.

Naive Bayesian [13] is one of a simple popular classifier algorithm based on probability. The core concept of this method is to find out the probability of the previously unseen instance belonging to each class, and then simply pick the most probable class. Naïve Bayesian classifiers are easy to implement, not sensitive to irrelevant features and widely used in making decisions about treatment processes.

Artificial Neural Network [13] is the most efficient classifier algorithm due to its result given more accuracy than the others. The fundamental processing of this algorithm is performed like human's brain. The element of a neural network is a neuron. This building block of human awareness encompasses a few general capabilities. Basically, a neuron receives inputs from any sources, combines them in some way, performs a generally nonlinear operation on the result, and then outputs the final result.

A. Description of dataset

This paper used the Health Facts database from Cerner Corporation, Kansas City, MO, which is a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States [14]. In dataset each record consists of 55 features. The features included are as follows : encounter identifier (ID), patient number , race, age, weight , admission type, discharge disposition, admission source, time in hospital, payer code, medical specialty, number of lab, number of procedures, number of medications, number of outpatient visits, number of emergency visits, number of inpatient visits, diagnosis 1, diagnosis 2, diagnosis 3, number of diagnoses, glucose serum test, A1c test result, change of medications, diabetes medications, 24 features for medications and readmitted.

B. Dose Titration Identification System

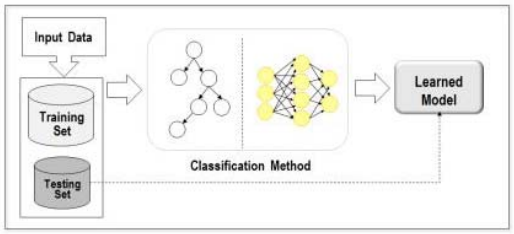


Fig. 2. Classification Model.

As shown in Figure 2, this paper propose to generate diabetes dose titration identification model by applying three classification algorithms; Decision Tree, Naïve Bayes and Artificial Neural Network. The experiment was initiated by data selection. Then selected data was separated into two sets as Independent Dose Titration Model (IDT) and Historical Dose Titration Model (HDT) prior than performing Data Mining process. Consequently raw data of each set was preprocessed and create identification model by applying three algorithms as above mentioned. Finally this paper uses the comparison of accuracy and ROC Curve results as Model Evaluation.

Data in this paper was extracted by based on Health Facts database in 130 hospitals through the United States in year 1999-2008. Data was covering clinical records from over 100,000 individual encounters. In data selection processing of type-2 diabetes inpatients for dose titration identified model, patients age 40 years old and over were selected as the target group. Drug categories were distributed in 4 groups; Metformin, Sulfonylurea, Thiazolidinedione and Insulin, subsequently classified into two aspects; Independent Dose Titration Model (IDT) and Historical Dose Titration Model (HDT) prior than performing Data Mining process.

For the first aspect; Independent Dose Titration Model (IDT), initial raw data was preprocessed, transformed some

attributes in order to resize data and reduce processing time purpose. The complete preprocessed data supplies in this paper comprise of six attributes and four drugs categories as per previously mentioned; Metformin, Sulfonylurea, Thiazolidinedione and Insulin. These dataset was used to create Independent Dose Titration Model as per Table 1

TABLE I. FEATURES OF INDEPENDENT DOSE TITRATION MODEL (IDT).

Feature name	Description and values
gender	Values: male, female
age	Grouped in 10-year intervals: 40, 50, …, 100
diag_1	The primary diagnosis (coded as first three digits of ICD9)
max_glu_serum	Indicates the range of the result Values: ">200", ">300" and "normal
A1c result	Indicates the range of the result , Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%
Met	Grouped in Metformin medical. Values : Up, Steady and Down
Sul	Grouped in Sulfonylurea medical. Values : Up, Steady and Down
TZD	Grouped in Thaizolidione medical. Values : Up, Steady and Down
Ins	Grouped in Insulin medical. Values : Up, Steady and Down

The second aspect; Historical Dose Titration Model (HDT) was similar to the first one, but data was more complicate in details than IDT. The historical record of dose titration was added by two previous times retrospectively. This dataset comprise of 17 attributes as per Table 2

TABLE II. FEATURE OF HISTORICAL DOSE TITRATION MODEL (HDT).

Feature name	Description and values
gender	Values: male, female
age	Grouped in 10-year intervals: 40, 50, …, 100
diag_1	The primary diagnosis (coded as first three digits of ICD9)
max_glu_serum	Indicates the range of the result Values: ">200", ">300" and "normal
A1c result	Indicates the range of the result. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%
Met02	Metformin value in retrospect to 2 times from recent therapy record. Values : Up, Steady and Down
Met01	Metformin value in retrospect to previous time from recent therapy record. Values : Up, Steady and Down
Met	Metformin value in recent therapy record. Values : Up, Steady and Down
Sul02	Sulfonylurea value in retrospect to 2 times from recent therapy record. Values : Up, Steady and Down
Sul01	Sulfonylurea value in retrospect to previous time from recent therapy record. Values : Up, Steady and Down
Sul	Sulfonylurea value in recent therapy record. Values : Up, Steady and Down
TZD02	Thaizolidione value in retrospect to 2 times from recent therapy record. Values : Up, Steady and Down
TZD01	Thaizolidione value in retrospect to previous time from recent therapy record. Values : Up, Steady and Down
TZD	Thaizolidione value in recent therapy record. Values : Up, Steady and Down Sulfonylurea
Ins02	Insulin value in retrospect to 2 times from recent therapy record. Values : Up, Steady and Down
Ins01	Insulin value in retrospect to previous time from recent therapy record. Values : Up, Steady and Down
Ins	Insulin value in recent therapy record. Values : Up, Steady and Down

C. Evaluation

This paper used the 10-fold cross validation method [15]. The 10-fold cross-validation is the measure the performance of the model to predict the sample. The base of this technique is re-sampling and takes one-tenth of data from data set as test data and test results from the identification of the classification model which is constructed by the rest of data.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In accordance with the knowledge discovery approaches discussed in the previous section, this section presents the results of the experiments into two parts: the experimental results of classification system.

A. Dose Titration Identification System

To build up the Dose Titration Predictive Model, this paper is proceeded by 3 selecting classifier algorithms; Decision Tree, Naïve Bayes and Artificial Neural Network. The related factors that be considered in comparison are the Accuracy and Receiver Operating Characteristic (ROC) Curve ratio. The results are satisfactory as reported in Table III and Table IV.

TABLE III. ACCURACY OF DECISION TREE, NAÏVE BAYES AND ARTIFICIAL NEURAL NETWORK.

Model Type	Medical	Accuracy		
		Decision Tree	*Naïve Bayes*	*Artificial Neural Network*
Independent Dose Titration Model (IDT)	Metformin	**0.62**	0.42	0.52
	Sulfonylurea	**0.51**	0.40	0.43
	Thiazolidinedione	**0.75**	0.45	0.62
	Insulin	**0.41**	0.38	0.38
Historical Dose Titration Model (HDT)	Metformin	0.56	0.44	**0.70**
	Sulfonylurea	0.43	0.41	**0.55**
	Thiazolidinedione	0.67	0.48	**0.84**
	Insulin	0.40	0.40	**0.43**

The comparison of the accuracy results obtained from the experiments is shown in Table III. The result operated by Decision Tree algorithms is provided the most satisfied accuracy in the range of 0.41-0.75 for Independent Dos Titration data (IDT). Meanwhile, the Artificial Neural Network algorithms provide the most satisfied accuracy in the range of 0.43-0.84 for Historical Dos Titration data (HDT).

Finally, Table IV present ROC Curve results obtained from the experiments. The ROC Curve results operated by Decision Tree algorithms is provided the most satisfied accuracy in the range of 0.60- 0.88 for Independent Dos Titration data (IDT). Meanwhile Artificial Neural Network

BIOGRAPHY

NAME	Miss Ratchanee Kaewthai
DATE OF BIRTH	16 April 1978
PLACE OF BIRTH	Lopburi, Thailand
INSTITUTIONS ATTENDED	Rajamangala University of Technology Suvarnabhumi, 2012-2014 Bachelor of Business Administration (Information Technology) Mahidol University, 2014-2015 Master of Science (Information Technology Management)
RESEARCH GRANTS	Grant to Support Graduate Students in Academic Presentations in Thailand Academic Year 2015
HOME ADDRESS	64/594 Moo 3 Monwadee Village, soi 2, Bangkurat, Bangbuathong, Nonthaburi, Thailand, 72160, Tel. 085-119-0994 E-mail : ratcha_ff@hotmail.com
PUBLICATION / PRESENTATION	Ratchanee K., Sotarat T. and Supaporn K., (2015), Diabetes Dose Titration Identification Model, The Proceedings of the 8th Biomedical Engineering International Conference (BMEiCON 2015), ISBN: 978-1-4673-9157-3.