Gülay ÇİÇEK

APPLYING DATA MINING TECHNIQUES TO IMPLEMENT THE CLINICAL GUIDELINES FOR THE MANAGEMENT OF PATIENTS WITH TYPE 2 DIABETES:MEDICATION DOSE ADJUSTMENTS

M.S.
2014

FATİH UNIVERSITY
The Graduate School of Sciences and Engineering

Master of Science in
Computer Engineering

APPLYING DATA MINING TECHNIQUES TO IMPLEMENT THE CLINICAL GUIDELINES FOR THE MANAGEMENT OF THE PATIENTS WITH TYPE 2 DIABETES: MEDICATION DOSE ADJUSTMENTS

by

Gülay ÇİÇEK

June 2014

APPLYING DATA MINIG TECHNIQUES TO IMPLEMENT THE CLINICAL GUIDELINES FOR THE MANAGEMENT OF THE PATIENTS WİTH TYPE 2 DIABETES: MEDICATION DOSE ADJUSTMENTS

by

Gülay ÇİÇEK

A thesis submitted to

the Graduate School of Sciences and Engineering

of

Fatih University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

June 2014
Istanbul, Turkey

APPLYING DATA MINING TECHNIQUES TO IMPLEMENT THE CLINICAL GUIDELINES FOR THE MANAGEMENT OF THE PATIENTS WITH TYPE 2 DIABETES: MEDICATION DOSE ADJUSTMENTS

Gülay ÇİÇEK

M.S. Thesis - Computer Engineering
June 2014

Thesis Supervisor: Assist. Prof. Dr. Kadir TUFAN

Co-Supervisor: Halil K. EROL, M.D.
Assist. Prof. of Internal Medicine

ABSTRACT

Diabetes is an essential community health problem which is approaching epidemic proportions globally. The prevalence of chronic, non-infectious disease is increasing at alarming rates worldwide. Every year 18 million people die from cardiovascular illness, for which diabetes mellitus and hypertension are the main predisposing elements. High or low blood sugar can result in numerous health complications. Many patients with type 2 diabetes meet with their health care provider every three four months; blood sugar levels and insulin dosing are evaluated at these visits, assisting the patients in diabetes management. Our aim is to create a tool that could be used to adjust medication dosage. A Medication dosage algorithm was developed using data mining techniques. Input parameters consisted of age, gender, alanine aminotransferase, aspartate transaminase, glycosylated hemoglobin (HbA1c), HDL cholesterol, creatinine, LDL cholesterol, microalbuminuria, triglyceride and urea. The dosage of the drug Metformin was adjusted according to the results of the algorithm.

Key words: Diabetes Mellitus, medication dose adjustment, data mining, classification

İLAÇ DOZLARININ AYARLANMASI: TİP 2 DİYABET HASTALARININ TEDAVİSİNDE KLİNİK KLAVUZLARIN GEREĞİNİ YERİNE GETİRMEK İÇİN VERİ MADENCİLİĞİ TEKNİKLERİNİN UYGULANMASI

Gülay ÇİÇEK

Yüksek Lisans Tezi – Bilgisayar Mühendisliği
Haziran 2014

Tez Danışmanı: Yrd. Doç. Dr. Kadir TUFAN

Eş Danışman: Yrd. Doç. Dr. H. Kutlu EROL

ÖZ

Diyabet, küresel salgın boyutlarına ulaşan önemli bir halk sağlığı sorunudur. Kronik hastalıkların yaygınlığı, endişe verici bir oranda artmaktadır. Diyabet ve hipertansiyon kalp ve damar hastalıklarının oluşmasına neden olmaktadır. Her yıl kalp ve damar hastalıklarından 18 milyon insan hayatını yitimektedir. Kanınızda şeker miktarının az veya çok olması, sağlık problemleri yaşamanıza sebep olabilir. Bundan dolayı şeker hastaları her üç ayda bir sağlık kuruluşuna gitmelidir. Kan şekerinin belli bir düzeyde tutmak için ilaç ve insulin dozları bu ziyaret sırasında tekrar gözden geçirilir ve uygun bir şekilde ayarlanır. Bizim amacımız doktorlara ilaç dozu ayarlarken onlara yardımcı olabilecek bir araç oluşturmak. Girdi parametreleri şunlardır : Cinsiyet , yaş, alanin aminotransferaz, HDL kolesterol, kreatinin, LDL kolesterol, mikroalbuminüri, trigliserid ve üre. Bu çalışmada metformin isimli ilaç için doz ayarlayamayı planladık.

Anahtar Kelimeler: Şeker Hastalığı, ilaç doz ayarlaması, veri madenciliği, sınıflama

To my parents, brother and sisters

ACKNOWLEDGEMENT

I would like to express my gratitude to my Allah which was a major source of strength when I worked on my thesis research.

A special thank you goes to my wonderful parent especially my father and to all of those who inspired, encouraged and fully supported me in during the writing of the thesis.

I want to thank Assist. Prof. Dr. Kadir TUFAN for his lead and insight throughout the research of thesis.

I am heartily thankful to my advisor, Dr. Halil K. EROL, for his help during the collection of the data. I am extremely thankful to him for sparing his valuable time and going through my thesis and making useful corrections and suggestions. Without his constant support, it would have been extremely difficult to finish this work.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

LIST OF SYMBOLS AND ABBREVIATIONS

SYMBOL/ABBREVIATION

ARFF Attribute Relation File Format

ANN Artificial Neural Network

DM Data Mining

FN False Negative

FP False Positive

ID3 Iterative Dichotomiser

ID3 Knowledge Discovery Databases

TN True Negative

TN True Positive

WEKA Waikato Environment for Knowledge Analysis

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Diabetes is a significant public health challenge that has reached epidemic proportions in recent years (Syed Amin Tabish, 2007). The overall prevalence on non-communicable disease has increased at a tremendous rate and as many as 18 million people succumb to cardiovascular disease each year; diabetes and hypertension are major risk factors for morbidity and mortality associated with heart disease. Current estimates suggest that as many as 312 million adults are obese while 1.7 billion are significantly overweight throughout the world. Furthermore, 155 million children may be overweight or obese. These statistics suggest that a worldwide diabetes epidemic is already underway. The International Diabetes Federation estimates that the relative prevalence of the disease was 8% in 2007 and will reach 7.3% by 2025 (Syed Amin Tabish, 2007). A total of 246 million people are affected by diabetes (46% of whom are between 40 and 59 years of age) and as many as 380 million may have active diabetes by 2025.

There are widespread concerns regarding this increase in type 2 diabetes. Many of people with diabetes in developed countries are above the retirement age. In those countries, many people are diagnosed with diabetes during middle age, between 30 and 65 years, which is the most productive phase of life.

According to estimation of International Diabetes Federation, every year 7 million human beings suffer from diabetes. Incidence of type 2 diabetes mellitus has increased dramatically. Lifestyle changes have been demonstrated to produce significant improvements in overall health among diabets patients.

Many patients with type 2 diabetes meet with health care providers every 3-4 months to evaluate medication usage, blood sugar, and insulin dosing. The primary treatment for type 2 diabetes is weight reduction. Physicians prescribe dietary restriction and exercise to patients with type 2 diabetes. When theses measure are unsuccessful for controlling high blood sugar, oral medication are used. Before adjusting oral medication, the physician examines some biochemical characteristics of diabetic patient such as blood and urine tests. As a result of these tests physicians give appropriate drugs and adjust dosages. Blood and urine analysis by experts is very time consuming and there is a shortage of experts possessing knowledge on the analysis of diabetic data. Methods to automate the interpretation of diabetic data by minimizing human efforts are critical for adjusting medication dosages. Data mining can be a solution to this problem by producing rules from those enormous datasets which can be used in analyzing diabetic data. Data mining is the practice of applying computation science to large data sets and extracting statistical inferences. Data mining has become a significant component of health care delivery and evaluation. Data mining includes classification, clustering, prediction and association of independent datasets. It also spans other disciplines like Data Warehousing, Statistics, Machine Learning and Artificial Intelligence (Farmer, 2007).

Data mining is estimated to be one of the most revolutionary developments of the century, according to the online technology magazine ZDNET New (Hornikova, Durakbasa & Güclü, 2011). Actually, The Technology Review Ten, (2001) chose data mining as one of 10 emerging technologies that will change the world. The examination of information from customers, partners, and suppliers has become important to many companies (Asadullaev, 2010).

Data mining can be useful in healthcare. Healthcare organizations that effectively implement data mining are able to better meet long term needs (Benko and Wilson 2003). Data can be a tremendous asset for healthcare organizations; however datasets must be adapted for digital manipulation (Hemalatha & Megala, 2011).

To estimate the result of a specific dosage of a medication is one of the most interesting applications of data mining. A recently developed resource, the Knowledge Discovery in Databases (KDD), is a widely used web application that enables the medical investigator to determine medication dosage based on a large

number of historical cases stored in a digital archive. A number of different classification methods can be used to evaluate this type of data, including Support Vector Machine, Neural Networks, and Decision Trees.

1.2 STATEMENT OF THE PROBLEM

The International Diabetes Federation (IDF) reports that diabetes mellitus affects 370 million individuals worldwide, among whom are 186 million individuals who are unaware of their illness. These statistics represent an increase over the 366 million estimated cases in 2011. Up to 4.8 million people will die as a result of the complications of diabetes each year, with patients under the age of 60 accounting for 50% of deaths (Sicree, Shaw & Zimmet, 2011)

Diabetes has been traditionally characterized as a disease of affluent countries. However, 60% of diabetes patients reside in low and middle income communities. Approximately 3.6 billion adults reside in high-risk low- and middle- income communities as of 2011. By comparison, 75 million cases of adult diabetes originate from high-income countries (IDF, 2011).

Type 2 diabetes accounts for 85% of diabetes cases worldwide. Many patients with type 2 diabetes meet with health care providers every 3-4 months to evaluate medication usage, blood sugar, and insulin dosing.

Our aim is to create a tool that could be used in adjusting dosage of medication. Medication dosage is adjusted using data mining techniques.

1.3 OBJECTIVES

1.3.1 General

The general objective of this study is to adjust dosage of medications by applying data mining techniques. To applying data mining technique, patient records need to be provided. Datasets were created by using patient records in Fatih University School of Medicine.

1.3.2 Specific

- To identify key features from datasets.
- To identify and select attributes that are more relevant in relation to diabetes
- To compare with Decision tree and Neural network classifiers in predicting dosage
- To interpret and examine the results of the selected model with the help of domain expert.

1.4 RESEARCH METHODOLOGY

In this study, the aim is to develop an estimation model that can predict the dosage of certain drugs based on the given datasets. The overall approach involved the use of the Knowledge Discovery in Database (KDD) methodology, the aim of which is to extract specific knowledge from data in the context of large databases.

1.4.1 To Understand the Application Domain

To define the problem and specify medical goals, I have worked closely with Dr. Erol, who is an internist at Fatih University School of Medicine. The discussions with this specialist have helped me clarify specific challenges and become familiar with the current solutions to those problems. Afterwards, a literature review was performed and relevant data mining techniques have been reviewed. A key sub goal in this technique is to determine the data mining success criteria. Our aim is to translate medical goals into data mining goals.

1.4.1.1 Selecting and Creating the Target Dataset

We used data which were collected by Fatih University School of Medicine and include patients with type 2 diabetes. There were 63 patient records in the dataset. To identify the most important features while creating the training data, I took advice from the medical expert. My medical mentor selected the most important eleven features as inclusion criteria in this dataset, which is called the “training data”.

1.4.1.2 Preprocessing and Cleansing

Data cleaning is the step where noise and irrelevant data are removed from the large data set (Garhwal, 2014). The results of the study are ultimately dependent upon the quality of the data input. In a given dataset, there can be duplicate records, unnecessary fields, and missing values. To increase the quality of the selected data, these problems must be addressed. In our dataset, we have some missing values and duplicate records. Missing values were replaced with the most probable value as determined by regression. To solve duplicate records in dataset, all records were reviewed and duplicate records were manually eliminated.

1.4.1.3 Data Transformation

The goal of data transformation is to reduce the number of effective variables to include only the most useful attributes which can address the specific goals of the task. Data is converted to the most appropriate format for data mining applications. In that case, after taking advice from the medical expert, a few transformations were required to make the data more suitable for data mining algorithms. In this study, we used two transformation methods which are called discretization and attribute selection.

To reduce the number of attributes in a dataset, attribute selection was applied. Attribute selection was needed to decrease the number of features as a classification algorithm to be examined and reduce errors resulting from inconsequential attributes. We have used the ranker search method to select the most appropriate attributes from 11 features that were available in the dataset.

Discretization was applied for converting continuous values variables to discrete values. This data transformation method was applied to reduce the number of distinct values of continuous variables by allowing a limited number of labels to demonstrate the original variables.

1.4.1.4 Selecting Suitable Data Mining Task

In this part, our aim is to determine what type of data mining modeling should be used. Choosing a suitable model based on the purpose of the study, and we aimed to adjust drug dosages using data mining techniques. There are different types of

algorithms for creating these models, such as Neural Network, Decision Tree, Support Vector Machine, and Bayesian Classification. To select the best model, we looked the performance of each algorithm.

1.4.1.5 Employing Data Mining Algorithm

Classification algorithm experiments were planned and conducted on a full training dataset including 63 instances. In all of the experiments two scenarios were considered, one containing all 11 attributes and the other containing only 8 selected attributes. A method of 10-fold cross validation was adopted for conducting random sampling of the training and test data sets.

1.4.1.6 Evaluation

In this step, models that are created will be evaluated according to their performance. The performance of individual models was evaluated using a series of algorithms including classification accuracy, ROC curve analysis, and a confusion matrix.

1.4.2 Tools

We need to use data mining algorithms for adjusting the drug dosage. Weka workbench, a collection of machine learning algorithms for data mining tasks, was used to determine drug dosage from the available datasets. In addition to that, there are some other reasons to why Weka was used. Weka is an open source application, which means that it is completely free to use and more significantly it is maintainable and alterable and contains code that is not dependent upon the commitment or longevity of any particular institution or company. Weka is run entirely in Java and can therefore be used on almost any hardware platform.

1.4.2.1 Weka Machine Learning Software

Weka (Waikato Environment for Knowledge Analysis, University of Waikato, New Zealand) is a widely used machine learning application and a collection of state-of-the art machine learning algorithms and data preprocessing tools for use in Java.

Weka is provided free of cost to the public under the GNU General Public License (Wikipedia, 2014).

Weka includes virtually all the popular data processing algorithms and is designed to allow the user to rapidly evaluate existing methodology on novel datasets in an interchangeable manner. The software provides extensive support throughout the experimental data mining process, including preparation of the input data, statistical evaluation of the learning schemes, and visualization of the input data. Weka includes both a wide range of learning algorithms and preprocessing tools in a diverse and comprehensive toolkit accessible through a common interface, allowing users to compare various methods and identify those that are most applicable to the given task (Chimieski & Faguned, 2013)

1.5 SIGNIFICANCE OF THE STUDY

The outcome of our study will reduce the number of medical errors by informing physicians in their determination of dosages of certain medications.

1.6 ORGANIZATION OF THE THESIS

The present thesis is divided into 6 sections. Chapter 1 provides an introductory explanation of the study design and goals. It also explains the objectives of the study, the research methodology, significance and scope of the study. Chapter 2 summarizes the relevant literature on data mining. Chapter 3 describes different algorithms and performance measures that are used during model and performance evaluation. Chapter 4 is concerned with understanding the implementation domain and the dataset selected for the study. Additionally, medical aims, data mining aims and dataset descriptions are offered in this section. Chapter 5 describes how the data is set up for a data mining task and the conversions that are executed on the data are presented. Chapter 6 deals with the empirical component of the study. In this chapter, empirical parts that are executed on the preprocessed data are discussed with their corresponding interpretations. Beside the point, the performance of each model created using different algorithms is presented. Chapter 7 of this thesis includes a summary of our study. The final Chapter sums up the methods that will be used, shows the test results and draws a general conclusion.

CHAPTER 2

2.1 OVERVIEW OF DATA MINING

With the increased popularity of data mining applications, an enormous amount of extant data is available for conversion into formats suitable to data mining processes. The overall goal of data mining is to reveal hidden relationships between data within large datasets. The term “knowledge discovery in databases” is generally synonymous with data mining. A number of additional terms have been used to describe data mining applications including information discovery, exploratory data analysis, information harvesting, knowledge extraction, and unsupervised recognition (Zhang & Hongwen, 2011).

Over the past five years, the Knowledge Discovery process has developed into a complex, multi-step process, with data mining being just one of these steps. Knowledge Discovery in Databases is a process that aims to identify useful information hidden within large databases. Widely used data mining techniques include data cleansing, data preparation and selection, incorporation of prior independent knowledge, and interpretation of the observed results. For extracting the information and pattern derived by the Knowledge Discovery in Databases, It is needed to use one of data mining algorithms. (Fayyad, Piatetsky & Smyth, 1996).

2.2 DEVELOPMENTAL HISTORY OF DATA MINING

While the term “data mining” first appeared in the 1990's, the process itself is the outcome of a long evolution. The storage of large amounts of data in computer systems for business purposes facilitated early efforts in data mining, and the field has advanced alongside technological changes in processing power and software. The mass storage of data on computers, disks, and tapes began in the 1960s and coincided with the introduction of relational databases and structured query languages. The use of

structured query languages enables users to evaluate stored data more effectively and dynamically (Berson, Smith, & Thearling, 2000).

The use of data in the 1960s differed greatly from modern practices. Business data mining has powerful applications, including the prediction of future financial conditions. The widespread growth of data collection necessitates the development of methods which can extract useful information. In particular, data mining technology aims to develop methods that will improve statistical evaluation, artificial intelligence and machine learning (Jibon, 2011).

2.3 KNOWLEDGE DISCOVERY IN DATABASE AND DATA MINING

Knowledge Discovery in Databases (KDD) is an automated tool for modeling relationships in large datasets. KDD is a process of identifying novel and useful patterns within large sets of complex data.

KDD has evolved through research on pattern recognition, database setup, statistics, machine learning, statistical inference, and artificial intelligence (AI). The aim of KDD is the extraction of high level information from large low-level datasets. For example; KDD addresses the means in which data are protected and accessed, how applied algorithms can be scaled to big datasets still run efficiently, how result can be interpreted and how the overall human- machine interaction can be modeled and supported usefully (Berry, Linoff & 2004)

Data mining (DM) is the core of the KDD process, including the use of inference algorithms to explore the data and the development of previously unknown data patterns. Models which explain complex phenomena can be generated by analysis and prediction methods.

2.4 DATA MINING TASKS

Data mining tasks may be semi-automated or fully automated analyses of large datasets which extract previously unrecognized patterns using data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining) (Wikipedia, 2014). Data mining applications can be either descriptive or predictive in nature.

A predictive model makes projects about future values of data based on known data. Predictive modeling is used to generate information about future trends. Companies use predictive model for identifying when the best time to run an advert for their production on television is using predictive modeling that takes into account their target demographic. Life assurance corporations utilize predictive modeling taking into account someone's age and way of life to identify what their premium would be. Descriptive models identify patterns or relationships within datasets, analyzing past events to generate insight into future events. Descriptive models aim to characterize the specific relationships that have determined success or failure in historical datasets. Descriptive analysis is widely used by consumer-driven organizations to assist in targeted marketing and advertisements. This type of analysis is used in sales, marketing, operations, and financial decision making (Dunham, 2004). The essential difference between predictive and descriptive methods is that descriptive models explore the relationships within historical datasets, while predictive models aim to project future events. Predictive models are necessary to project future outcomes and evaluate progress towards these goals.

Methodology associated with predictive models includes classification, prediction, regression and time series analysis. Clustering summarization, association rules, and sequence discovery are typically associated with descriptive methods.

2.4.1 Classification

Classification processes categorize data according to pre-defined criteria. The term supervised learning has also been used to describe these processes (Jibon, 2002). Examples of classification tasks include the following:

- Identification of fraudulent credit card transactions
- Assessing credit risk of a mortgage application
- Diagnosis of disease
- Verifying the authenticity of legal documents, such as a will
- Identification of terrorist threats based on financial transactions or personal behavior

Classification involves the prediction of a future result based on a given input. Classification algorithms process training data sets that include attributes associated with particular outcomes, generally known as the goal or estimation attributes. Classification algorithms identify associations within the dataset that enable accurate estimation of outcomes. Subsequently, a prediction dataset (estimation set) including the same attributes with the exception of the prediction attribute is used as an input for the algorithm. The classification algorithm analyzes the input and generates a prediction estimate. The accuracy of the prediction based on known outcomes is used to determine the efficacy of the algorithm. For example, a classification algorithm may be trained with a given medical dataset. A prediction set containing the same attributes can be used to identify the likelihood of heart problems (Viana, 2007). Table 2.1 is an example of the use of training and prediction datasets:

Table 2.1 Training Set.

Age	**Heart Rate**	**Blood pressure**	**Heart problem**
60	72	150/70	Yes
36	82	112/76	No
70	63	108/165	No

Table 2.2 Prediction Set.

Age	**Heart Rate**	**Blood pressure**	**Heart problem**
42	97	148/89	?
64	57	106/63	?
83	76	150/65	?

Algorithm: If (age=65 and heart rate>70) or (age>60 and blood pressure>140/70) then heart problem='yes'

Estimation (prediction) rules are given in the form of if-then statements, with the if statement consisting of specific logical conditions and the rule consequence (then) statement generating a prediction attribute that satisfies the antecedent if logical statement. The use of conjunction statements allows individual logical rules to be specifically stated and defined in relation to each other. Alternate classification

algorithms include decision trees, neural networks, and Bayesian classification (Leanordo, 2007).

2.4.1.1 Decision Tree

A decision tree is a classification algorithm with recursive or iterative partition of the instance space; it consists of a series of nodes forming a rooted tree in which the root has no incoming edge. Nodes with outgoing edges are frequently referred to as internal or test nodes. Terminal nodes are referred to as “leaves” or decision nodes. Internal nodes in a decision tree are divided into two or more subspaces as a result of a specific discrete function of the input attribute. In the most basic sense, each test evaluates a single attribute, and the instance space is separated according to attribute values. A range is often applied for numeric attributes. Individual leaves on a decision tree are directed at one class and indicate the most appropriate target value. Circles represent internal nodes while leaves are indicated by triangles. Multiple branches may be drawn from each internal node. Each node and its branches correspond to a certain value range. This range specifies a partition within the possible values of a given attribute.

Individual instances are categorized by navigation from the root node to the terminal leaf, applying test characteristics at each intermediate node. A decision tree contains both nominal (categorical) and numeric features. Nodes are characterized by the attribute tested and branches are labeled according to specific value partitions. Similar data mining algorithms are widely used in medicine, molecular biology, manufacturing, and financial analysis (Han & Micheline, 2003).

2.4.1.2 Artificial Neural Networks

Artificial Neural Networks (ANN), or more generally, neural networks (NN), are a computational emulation of biological neural networks. Neural network architecture includes an interconnected group of artificial neurons that process information inputs by a connection-mediated algorithm. In many applications, the ANN is a dynamic and adaptive system that changes in structure as a result of external or internal data inputs received during a training phase (Jibon, 2011).

Neural networks are generally used because of the convergence of various factors. One of the advantages of using a neural network is this technique is quite robust with respect to noisy data. Another advantage to using neural networking is learning. Using a neural network, we can solve many problems without finding and describing a specific method of such problem solving, without building algorithms by ourselves, and without developing programs. We can use a neural network, which first learns the results of a solved problem and uses this to solve many other problems. It is really a very comfortable and efficient way of problem solving!

2.5 APPLICATION OF DATA MINING IN THE HEALTH SECTOR

Data mining has been applied to many fields and has proven to be particularly useful in healthcare applications. The results of data mining can benefit patients, healthcare providers, insurers, and regulators. Healthcare organizations use data mining to improve customer relationship and management decisions. Physicians use data to make evidence-based treatment decisions, and patients receive better care as a result. Data generated in the healthcare setting can be very complex and may involve extremely large volumes of information, making analysis of healthcare data exceptionally challenging. Data mining methodology and technology can transform large collections of data into useful knowledge that informs healthcare decisions.

2.6 DIABETES MELLITUS

Diabetes is a potentially life-threatening disease that occurs as the result of insufficient insulin production. The hormone insulin regulates the uptake of glucose by all cells in the body. High blood sugar levels resulting from insufficient insulin production can have many damaging effects of bodily tissues. Elevated blood glucose or hyperglycemia occurs in patients with poorly controlled diabetes. Chronically elevated blood glucose can damage the kidneys, nerves, blood vessels, and feet. In addition to that, diabetes can also cause high blood pressure and hardening of the arteries or arteriosclerosis. Diabetes occurs when your body does not make enough insulin or cannot use the insulin it makes. Insulin is a hormone that controls the amount of sugar (glucose) in the blood. A high blood sugar level can cause problems in many parts of the body. These can bring on blood vessel disease and heart problems (National Kidney Disease, 2014).

Symptoms of diabetes vary depending on the specific type of diabetes. A classic symptom of diabetes is frequent urination and excessive thirst. However, some forms of diabetes such as type 2 diabetes do not have noticeable symptoms in the early stages. Patients may not experience symptoms for many years prior to diagnosis.

As many as 347 million people worldwide suffer from diabetes. An estimated 3.4 million individuals died in 2004 as a result of uncontrolled hyperglycemia. More than 80% of deaths associated with diabetes occur in low and middle-income countries. The WHO predicts that diabetes will be the 7th leading cause of death worldwide by the year 2030.

2.6.1 Major Types of Diabetes

There are four types of diabetes, but in this study the researcher addresses two types that are most common. Two types of diabetes are explained below.

2.6.1.1 Type 1 Diabetes

Individuals with type I diabetes do not produce insulin as a result of childhood autoimmunity. Generally, this disease is seen in young adults or children, but it can occur at any age. Patients with type 1 diabetes must manage their blood glucose using insulin shots or other insulin administration devices. Approximately 7-12% of all cases of diabetes are type 1.

2.6.1.2 Type 2 Diabetes

Type 2 diabetes is caused by insufficient production or utilization of insulin. To preventing this type of diabetes, physicians prescribe dietary restrictions to patients. In addition to that, patients are often advised to exercise. Usually, this type of disease occurs in individuals over 40, but it can be seen earlier. Patients with type 2 diabetes may require insulin shots; however the general treatment involves oral medication, dietary restriction, and exercise. Type 2 diabetes is the most common form of diabetes.

CHAPTER 3

ALGORITHMS USED FOR MODEL BUILDING AND PERFORMANCE MEASURES

In this study, our aim is to estimate medication dosage for a given patient with type 2 diabetes. An attempt was made to construct a prediction model using Decision Tree and Neural Network methods. After building the models, performance of each of the models were evaluated, and their performances were compared to each other. In this part, the algorithms used to build the models and the matrices used for performance measures and comparisons are discussed in detail.

3.1 DECISION TREES

A decision tree resembles a flowchart, with each attribute tested at the branch points and the result of the test determining node assignment. There is a path from root to leaf. This path specifies classification rules. The top node in decision tree is called the root node.

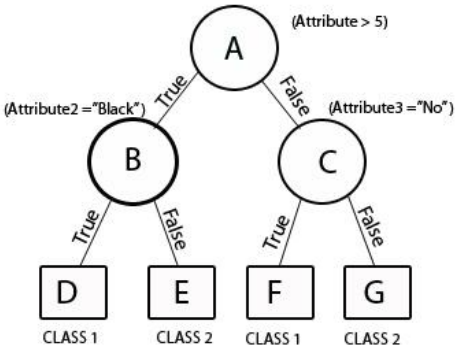

Figure 3.1 Simple Decision Tree.

Many human analysts find decision tree systems easy to understand. The discrete steps of classification within a decision tree occur quickly and simply. Decision tree algorithms may achieve high levels of accuracy; however this will be dependent upon the amount and quality of the data available. As in many areas, Decision tree algorithms are used such as medicine, production and manufacturing and so on.

3.1.1 J48 Classifier Algorithm

J.Ross Quinlan, a machine learning researcher, has developed an improved decision tree algorithm known as Iterative Dichotomiser (ID3). A later successor, C4.5, improved upon the Iterative Dichotmomiser (ID3).

This algorithm is implemented in a serial manner. A pruning method used in C4.5 replaces internal nodes with leaf nodes, reducing the overall error rate. The ID3 algorithm functions as a non-categorical feature. However the C4.5 algorithm can utilize both continuous and categorical features when generating a decision tree. Both ID3 and C4.5 utilize a gain ration to determine the optimal splitting attribute.

Entropy reduction is used to generate the optimal splitting attribute in C4.5

- Given probabilities p_1, p_2, .., p_s, whose sum is 1, Entropy is defined as:

$$H(p_1, p_2, .., p_s) = \sum_{i=1}^{s}(p_i \log(1/p_i)) \quad (3.1)$$

Given candidate split S, which partitions the training dataset T into several subsets, T1, T2,,,,,,, Tk the mean information requirement can then be calculated as the weighed sum of the entropies for the individual subsets, as follows:

$$Hs(T) = \sum_{i=1}^{k}(p_i \, Hs(T_i)) \quad (3.2)$$

Gain can subsequently be calculated using the following formula:

$$Gain(D, S) = H(D) - \sum_{i=1}^{s}(p(D_i) \, H(D_i)) \quad (3.3)$$

At each successive decision node, C4.5 determines the optimal split based on maximized information gain. J48 implements the C4.5 algorithm to generate both pruned and un-pruned C4.5 decision trees. Decision trees generated by J48 can be used for accurate classification.

J48 can utilize both discrete and continuous attributes as well as training data with missing attribute values and attributes with variable cost assignments.

The C4.5 algorithm for generating decision trees can be applied to classification problems. It advanced on the ID3 algorithm in several ways:

- **Missing data:** Missing data are disregarded during construction of the decision tree. Gain ratio is determined based solely of records that have a value for a given attribute. For classifying a record with an incomplete attribute value, the value for that item can be estimated based on what is known about the attribute values for the other records.
- **Continuous data:** The basic concept is to split the data into ranges based on the attribute values for that item in a training sample.
- **Pruning:** Pruning is a way of reducing the size of the decision tree. It will mitigate the accuracy of the training data, but enhances the accuracy given unseen data.
- **Rules:** C4.5 allows for classification via decision trees or regulations created from them.
- **Splitting:** The maximal gain ratio ensuring ah larger than average information gain is used for splitting.

Classification algorithms require a set of training data; each instance occurs as paired input and output objects. The classification algorithm builds classification criteria based on the training dataset that correctly classifies both training and test examples. Test inputs to the algorithm are used to generate output values (Kilanly, 2013).

3.2 NEURAL NETWORKS

Neural networks are interconnected nodes with directional links. Individual nodes represent processing units, and the links between nodes reflect causal links between these processes. Neural network systems were originally generated to mimic the neurophysiology of the human brain using simple computational elements (neurons) in a highly interconnected system (Kilanly, 2013).

Neural networks became popular in the 1980s because of a convergence of many factors. First, Neural Networks (NN) is more robust than Decision Trees because of weighted factors. Second, the NN improves performance by learning. Third, learning may continue even after a training set has been applied. Fourth, the use of NNs can be parallelized for better performance. Fifth, there is a low error rate and thus a high degree of accuracy once the appropriate training has been applied. Because of these factors, neural networks are, and will continue to be, popular tools for data mining.

Neural network architectures contain at least two layers, an input and an output layer. Additional layers can be inserted between the input and output. The input layer may contain multiple nodes, each of which represents a unique predictor variable. Nodes are the foundation of the neural network. All input nodes can be linked to other input nodes within the hidden layer, to other layers, or to the output. Multiple nodes in the output layer represent response variables. Each node receives a set of inputs which are multiplied by connection weight Wxy (e.g. the weight from node 1 to 3 is W13) and added together (Two Cross Corporation, 1999). The activation function is then applied to the output before passing the function to the next layer.

Activation function is applied as following

- Multiplies each input by its weight
- Applies an activation function to the sum of results.

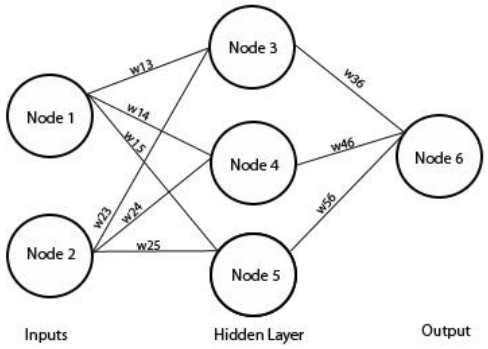

Figure 3.2 A Neural Network Architecture.

The individual nodes can be considered estimated variables (nodes 1-2, in this example) or as a combination of predictor variables (nodes 3-6). Node 6 is the non-linear combination of the values of node 1-2, resulting from the activation function using the summed hidden node values (Two Cross Corporation, 1999). In the absence of hidden layers, linear activation functions within the neural network are equivalent to linear regression; certain non-linear activation functions of neural nets are equivalent to logistic regression. The topology of a neural network includes a large number of nodes. The user or software parameters select the hidden nodes and layers, the activation function, and the parameters of the network design (Two Cross Corporation, 1999). It is used below formula for deciding number of hidden layer. Let's assume that A represent number of hidden layer.

A = (number of attributes +number of classes) / 2

3.2.1 Multilayer Perception

Multilayer perception (MLP) is a type of neural network that has become the most widely used in data mining. Multilayer perception includes multi-layer nodes arranged in a directed graph. All nodes in a layer are linked to a subsequent layer. Neurons, or nodes within layers, have non-linear function except for the input layer. A supervised training technique known as back propagation trains the network in multilayer perception.

The forward feed of the network propagates from the input to the output without iteration. A feed forward neural network does not include connections

between nodes in the same layer; output nodes in a specific layer are linked to specific input nodes in succeeding layers. This modular design is generally preferred, with nodes in a shared layer sharing some functionality or generating comparable abstractions (Che &Tzu, 2010).

3.3 PERFORMANCE MEASURES

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and another used to validate the model (Ferri, Orallo & Modroiu, 2007). The primary method of statistical cross-validation is the K-fold method. In K-fold cross-validation, data are randomly divided into mutually exclusive subsets or 'folds' D1, D2,……, Dk, each with similar size. Testing and training are implemented k times (Refaeilzadeh, Tang & Huan (2008).

There are two primary reasons to use cross validation statistics

- To generate algorithm performance data associated with a learned model based on existing data
- To compare performance among multiple algorithms using a given dataset

3.3.1 10-Fold Cross Validation

In 10 –fold CV each dataset is randomly separated into 10 reciprocally exclusive subsets of comparable size. The constructed model is trained and tested 10 times. The algorithm is trained in the first iteration and tested in the final iteration.

In order to perform of 10 fold CV, a set of general principles may be followed:

Pace 1: Dataset is divided into 10 equal parts. Each part is known as a fold.
Pace 2: Create a model using all records in each fold except one fold. The excepted fold is used for the testing purpose.
Pace 3: Step 2 is replicated 10 times and average accuracy is calculated.

3.3.2 Confusion Matrix

Many different mechanisms are used to measure performance in classification problems. If there are k classes, the size of the row and column of a confusion matrix is equal to k. Size of the table is k by k. If instance is positive and this is classified as positive; it is counted as TP; if it is classified as negative, it is counted as a false negative (FN). If the instance is negative and it is classified as negative, it is counted as a false negative (FN). If the instance is negative and it is classified as negative, it is counted as true negative (TN); positive instances are considered false positive (FP)". Accurate identification of true/false positive and true/false negative results is critical to determining the accuracy of the algorithm.

		Predicted class	
		C1	C2
Actual class	C1	true positive	false negative
	C2	false positive	true negative

Figure 3.3 A Simple Confusion Matrix.

Figure 3.3 shows a confusion matrix for binary classification. The numbers along the diagonal from upper-left to lower-right demonstrate the correct decision made, and the numbers outside this diagonal demonstrate the errors.

Other performance measures can also be performed, including specificity, recall, and F-measure.

The method to determine the accuracy of a given classifier is as follows:

Step 1: A= the sum number of true positive results and true negative results

Step 2: B= the sum number of true positive and negative results and the false positive and negative results

Step 3: Divide A by B

The accuracy of a classifier is defined as the proportion of test instances that are correctly classified.

$$\text{Accuracys} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (3.4)$$

3.4.3 Area under the ROC Curve

"A Receiver Operating Characteristics (ROC) curve is a techniques for visualizing, organizing and selecting classifiers based on their performance. In essence, it is another performance evaluation technique for classification models and also useful tool for comparing two on more classification models. ROC curves have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates of classifiers. The use of ROC analysis has been extended into visualizing and analyzing the behavior of diagnostic systems. Recently, the medical decision making community has developed an extensive on the use of ROC curves as one of the primary methods for diagnostic testing " (Olson and Delen, 2008).

"In order to plot an ROC curve for a given classification model, true positive (TP) rate is plotted on the Y axis and false positive (FP) rate is plotted on the X axis. Roc curve we start at the bottom left-hand corner (where the true positive rate and false positive rate are both 0), we check the actual class label of the tuple at the top of the list. If we have a true positive (that is a positive tuple that was correctly classified), then on the ROC curve, we have a false positive. On the ROC curve, we move right plot a point" (Olson and Delen, 2008).

Table 3.1 Performance measure of a ROC Area.

ROC Area	**Performance**
0.9 -1.0	Excellent(A)
0.8-0.9	Good (B)
0.7-0.8	Fair (C)
0.6- 0.7	Poor (D)
0.5- 0.6	Fair (F)

CHAPTER 4

BUSINESS AND DATA UNDERSTANDING

This study is based on the data gathered from Fatih University School of Medicine. To better understand the medical field, the researcher worked closely with Halil K Erol, M.D., who is an internist at Fatih University School of Medicine. In addition, real time observation of the business process was performed to gain insight into how the hospital functions. The first step was to observe the current procedures that are used to adjust the drug dosages. After the procedures were identified, the researcher continued to define problems that occurred during the medical consultation. Finally medical goals were determined and data mining goals were set to identify the data required for the study.

4.1 BUSINESS PROCESS DESCRIPTION

When the patient comes to the hospital, he/she directly requests for the service at the reception desk. After registration, the patient's medical chart is sent to the physician's computer. When the physician is ready to examine the patient, the patient enters to the examination room and the physician listens to his/her complaints and may request some laboratory tests. When the results are ready, they will be sent to the physician's computer from the laboratory directly. Next, the physician discusses the results with the patient and may prescribe drugs.

4.2 ADJUSTING MEDICATION DOSAGE

There are many methods that can be used to adjust drug dosage. Generally the adjusting of the drug dosage starts with identifying the patient's history and habits

and by performing a physical examination. Laboratory tests are used to adjust the drug dosage.

4.3 MEDICAL PROBLEM DEFINITION

4.3.1 Defining Medical Goals

The goals of the medical study were defined as follows:

- To create a tool that could be used in adjusting drug dosage.
- To make the interpretation procedure easier, more consistent, and efficient based on the rules that are formulated by the system.

4.3.2 Defining Data Mining Goals

The data mining goals are translations of the medical goals. Here the goals are set towards the technical part of the solution.

In order to solve the problems that exist in current system, the following medical goals were set:

- To create a tool that could be used in adjusting drug dosage.
- To make the interpretation procedure easier, more consistent, and efficient based on the rules that are formulated by the system.
- To specify key attributes or patterns from the dataset.
- To specify and select attributes that are important in relation to predictable state- adjusted dosage.
- To apply neural network, decision trees and Bayesian classifier to the analysis of the dataset

4.4 Selecting and Creating the Target Dataset

After establishing data mining aims and the project plan, the next step was selecting and creating a target suitable for the study. This includes identification of the available data and missing data containing all the attributes considered in the data mining process. Learning processes employed in data mining are defined by the test

input data; therefore precise definition of the test dataset is critical to the ultimate function of the application.

4.4.1 Data Understanding

Data from the patient's histories contain laboratory test results. This data is usually created and stored in the context of a medical decision, for the purpose of augmenting choices about further testing and/or treatment, and also for future access when required.

4.4.2 Description of the Dataset

Each recording in the dataset represents a single patient exam result gathered during medical consultation. The attributes in the dataset contain age, gender, alanine aminotransferase, aspartate transaminase, glycosylated hemoglobin (HbA1c), HDL cholesterol, creatinine, LDL cholesterol, microalbuminuria, triglyceride and urea. The attributes and their illustration are presented in Table 4.1

Table 4.1 The Attributes and their illustration.

No	Attribute	Description	Type
1	Age	Age of patient	Numeric
2	Gender	Gender of patient	Nominal
3	Alanine Aminotransferase	Liver function test	Numeric
4	Aspartate Transaminase	Liver function test	Numeric
5	Glycosylated Hemoglobin	Average glucose level	Numeric
6	HDL Cholesterol	Good cholesterol	Numeric
7	Creatinine	Kidney function test	Numeric
8	LDL cholesterol	Bad cholesterol	Numeric
9	Microalbuminuria	Kidney function test	Numeric
10	Triglyceride	Blood lipid level	Numeric
11	Urea	Kidney function test	Numeric

CHAPTER 5

DATA PREPROCESSING

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data because of their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results.

The presence of missing, noisy and inconsistent data is commonplace in large datasets. There can be missing value in a dataset for many reasons. Some attributes of interest may be missing or incomplete in the dataset while others may be missing simply because they were not considered important at the time of data entry.

Attribute or feature may not be recorded as a result off misunderstanding or malfunctioning equipment. Incoherent data may have been inadvertently discarded at any time in the history of the recorded documents (Han and Kamber, 2006).

Preprocessing techniques include data cleaning to configure missing values, noise removal, outlier processing, and the correction of obvious inconsistencies.

5.1 DATA FORMATS

Weka machine learning software was used in the present study, and it was necessary to convert the data input to a suitable format, the ARFF file. An ARFF (Attribute-Relation File Format) file is an ACII text file that describes a list of instances sharing a set of attributes (Witten & Frank 2005).

ARFF Files have two distinct components:

1. Metadata
   - Relation name
   - Domains and attributes
2. Data
   - Individuals rows or instances of a relation

5.2 DATA CLEANING

Data cleaning is the process of identifying, correcting, or removing inaccurate or incomplete records from a database.

5.2.1 Handling Noisy Data

Noise is defined in this case as random variance in a measure variable. There are many possible causes of noisy data, including human or electronic error. Errors can also be introduced during data transmission. Technological restrictions such as buffer size can corrupt data transfer and synchronization. One of the common types of error was spelling errors. For example, a word was misspelled “mala” several times. Researcher identified this and replaced the term with the correct spelling.

5.2.2 Handling Missing Values

A widely used method was applied to the dataset to determine the most probable replacement value in the case of missing data. This method relies on the extant attributes of a given instance to project likely values for missing data. This method finds the missing value and replaces the missing value with predicted ones.

5.2.3 Handling Inconsistent Data

Same data is presented in several ways. If these situations are not caught, when these situations are not caught they might cause wrong results, therefore these cases were handled to prevent misleading results. Gender has only two distinct values, male and female, although these values may be noted in a variety of formats. In some instances, the gender of the patient was saved using abbreviations (M and F). In

some other cases, these values were saved using the all form of the words (Male and Female).In this study, researcher face same problem. For preventing these inconsistencies this situation presented with the word ‘m’ were changed to ‘male’.

5.3 DATA TRANSFORMATION

Once the data has been assembled and major data problems are fixed, the data must still be converted for analysis. This includes adding derived fields to bring information to the surface. It may also include smoothing, aggregation, generalization, normalization, discretization, and attribute construction.

Data transformation methods are necessary to adapt data to formats conducive to data mining. Discretization was used to reduce distinct values of attributes and an attribute selection method was used to eliminate poorly relevant attributes.

5.3.1 Discretization

Age was broken into discrete values in a process known as discretization. Discretization is the process of converting continuous variables into discrete value where a limited number of labels are used to represent the original values. Discrete values have a limited number of intervals along a continuous spectrum while true continuous variables have an infinite number of intervals.

Olson and Delen (2008) proposed several justifications for the use of discrete intervals

- Discrete values are more appropriate for knowledge-level representation
- Discretization reduces and simplifies data
- Discrete attributes are easier to understand for both experts and novice users.

5.3.2 Attribute Selection

A second reduction technique applied to the dataset was attribute selection. When the database was created, medical experts selected the attributes that were included in the data set, but still that was not enough. Attribute selection eliminates redundant attributes. By reducing the number of attributes within the reported patterns, these patterns often become easier to understand.

To reduce the number of attributes, the researcher applied a ranker method using all attributes combined with some of the available attribute evaluators for that method. “Ranker” will give you a list of the attributes, ordered by their score according to the evaluator.

Table 5.1 The attributes and Ranked Value.

Parameter	Ranked Value
Glycosylated hemoglobin	0.703
Age	0.675
Creatinine	0.397
Triglyceride	0.318
LDL Cholesterol	0.315
HDL Cholesterol	0.311
Aspartate Transaminase	0.291
Alanine Aminotransferase	0.285
Urea	0.156
Microalbuminuria	0.102

CHAPTER 6

EXPERIMENTATION

As the aim of this study is to adjust medication dosage using data mining techniques, a classification technique was adopted to improve a predictive model. The models were built with two separate supervised machines, i.e. Decision Tree and Neural Network using Weka 3.6.4 machine learning software.

6.1 EXPERIMENTAL SETUP

For this study, experiments were handled and every experiments two scenarios were considered, one including all the 11 features and the other including 9 selected attributed. So with each experiment and eight different scenarios a total of eight models were built.

First Experiments were conducted on a full training dataset including 63 instances and 10-Fold Cross Validation was adopted for randomly sampling the training and test sets. While performing the experiments all parameters were set to their default setting for each algorithm except for J48 classifier where the parameter “Unpruned” which had a default value “False” was changed to “True” for the first experiment to observe the performance of J48 unpruned tree.

In this study, the performance of models were evaluated using standard metrics of accuracy, precision, recall and F-measure which were computed using predictive classification table, known as Confusion Matrix. For comparing the performance measures, Receiver Operating Characteristic was also used.

6.2 MODEL BUILDING USING J48 DECISION TREE

In J48 Decision Tree Classifier, Two experiments were conducted. We have designed two experiments for investigating:

- The effect of attribute selection on classification accuracy on unpruned J48 Decision Tree Classifiers.
- The effect of attribute selection on classification accuracy on pruned J48 Decision Tree Classifiers.

Experiment 1

The first experiment was designed to evaluate the performance of a J48 classifier unpruned tree in predicting dosage of medication and investigate the effect selection on the performance of the model. In this experiment two scenarios were considered, one including all 11 attributes and the other including the selected 9 attributes.

One first scenario the algorithm was run on a full training set including 560 (number of instances is increased because of imbalanced dataset) instances with 11 attributes.On the second scenario the algorithm was run on a cross validation set including 560 instances with selected 9 attributes.

Table 6.1 Confusion Matrixes for Experiment 1.

Model	0(Predicted)	500(Predicted)	1000 (Predicted)	1500 (Predicted)	Actual
J48 unprunned with all attributes	158	9	5	4	**0**
	2	140	1	17	**500**
	5	3	81	7	**1000**
	0	1	3	124	**1500**
J48 unprunned with selected attributes	158	9	5	4	**0**
	2	142	1	15	**500**
	5	3	79	9	**1000**
	0	2	4	122	**1500**

Table 6.2 Detailed Performance Measures for Experiment 1.

Model	Test Options	Accuracy	TP Rate	FP Rate	Precision	F-Measure	ROC Area
J48 unprunned with all attributes	**Cross Validation**	89. 82 %	0.89	0.03	0.90	0.89	0. 94
J48 unprunned with selected attributes	**Cross Validation**	89. 46 %	0.89	0.03	0.89	0.89	0.94

The model built with J48 unpruned tree with all attributes correctly classified 503 (89. 8214%) instances while 57 (10,1786%) of the instances were classified incorrectly. The overall accuracy rate of the model is highly successful.

The second model was built with J48 unpruned tree with selected 9 attributes correctly classified 501 (89,4643%) instances while 59 (%10,5357) of the instances were classified incorrectly. Like the experiment the overall accuracy rate of the model is highly successful.

Experiment 2

The second experiment was designed to investigate:

- The performance of a J48 classifier pruned tree in estimating dosage of medication.
- The effect of attribute selection on the performance of a J48 classifier pruned tree model.
- The effect of the tree pruning methods when building a J48 decision model

Similar to experiment 1, in this experiment two scenarios were considered, one containing all 11 attributes and other containing the selected 9 attributes. On the first scenario the algorithm was run on a full training set including 560 instance with all attributes.

On the second scenario the algorithm was run on a full training set including 560 instances with only 9 selected attributes.

Table 6.3 Confusion Matrixes for Experiment 2.

Model	0(Predicted)	500(Predicted)	1000 (Predicted)	1500 (Predicted)	Actual
J48 prunned with all attributes	158	9	5	4	**0**
	2	140	1	17	**500**
	5	3	81	7	**1000**
	0	1	3	124	**1500**
J48 prunned with selected attributes	158	9	5	4	**0**
	2	141	1	16	**500**
	5	3	78	10	**1000**
	0	2	4	122	**1500**

Table 6.4 Detailed Performance Measures for Experiment 2.

Model	Test Options	Accuracy	TP Rate	FP Rate	Precision	F-Measure	ROC Area
J48 prunned with all attributes	Cross Validation	87. 14 %	0.87	0.04	0.88	0.87	0. 94
J48 prunned with selected attributes	Cross Validation	87. 85 %	0.87	0.04	0.88	0.87	0. 94

6.3 MODEL BUILDING USING NEURAL NETWORK

This experiment was designed to explore the ability of Neural Network in estimating dosage of medication. From Neural Network Algorithms Multilayer perception was selected to conduct in experiment. In this experiment like in all of the experiment two scenario was conducted one containing all 11 attributes and other containing the selected 9 attributes.

Experiment 5

On the first scenario the algorithm was run on a **cross validation** set containing 560 instances with 11 attributes. On the second scenario the algorithm was run on a **cross validation** set containing 560 instances with 9 selected attributes.

Table 6.5 Confusion Matrixes for Experiment 3.

Model	0 (Predicted)	500 (Predicted)	1000 (Predicted)	1500 (Predicted)	Actual
Neural Network with all attributes	161	7	7	1	**0**
	2	131	0	27	**500**
	12	4	70	10	**1000**
	0	1	1	126	**1500**
Neural Network with selected attributes	164	5	7	0	**0**
	2	131	1	26	**500**
	12	4	71	9	**1000**
	0	1	1	126	**1500**

Table 6.6 Detailed Performance Measures for Experiment 3.

Model	Test Options	Accuracy	TP Rate	FP Rate	Precision	F-Measure	ROC Area
J48 prunned with all attributes	**Cross Validation**	87. 14 %	0.87	0.04	0.88	0.87	0. 94
J48 prunned with selected attributes	**Cross Validation**	87. 85 %	0.87	0.04	0.88	0.87	0. 94

The first Neural Network model built on all 11 attributes correctly classified 488 (87, 1429 %) instances while 72(12, 8571 %) of the instances were classified incorrectly. The overall accuracy rate of the model is very high. But it is not better compared to J48 pruned and unpruned model.

The second Neural Network model built on 8 selected attributes correctly classified 492 (87, 1229 %) instances while 62(12, 8771 %) of the instances were classified incorrectly.

6.4 DISCUSSION

Six experiment were conducted and classification performance has been compared in order to determine optimal statistical algorithms for predicting dosage of medication. The experiments were designed for four purposes; to investigate the effect of tree pruning when building a decision tree model, to observe how attribute selection affects the classification accuracy, to compare J48 Decision Tree classification algorithm and Neural Network.

CHAPTER 7

CONCLUSION

In this study, the aim was to design a predictive model for adjusting dosage of medication using data mining techniques. Data collected by Fatih University School of Medicine from the year including 63 instances was selected and preprocessed for this study. The models were built on the preprocessed Metformin Dataset with two different supervised machine learning algorithms i.e. J48 Classifier and Multilayer Perception using Weka 3.6.4 machine learning software.

The performance of the models were evaluated using the standard metrics of accuracy, precision, recall and F- measure. 10 – Fold Cross Validation was adopted for randomly sampling the training and test data samples. All fourteen models performed well in predicting dosage of medication. The most effective model to be J48 classifier implemented on selected attributes and all attributes with a classification accuracy of 91.76 %.

From a total of 11 attributes that were available, all of them are highly relevant in estimating medication dosage of medication from Metformin dataset were selected. Because when we remove two feature from dataset, we can acquire low performance.

This study showed that data mining techniques can be used efficiently to model and predict dosage of medication. The outcome of this study can be used as assistant tool by physicians.

REFERENCES

Asadullaev, S., " Data warehouse architectures and development strategy ", *Companion Guidebook*

Berry, M. J. A., Linoff, G., *Data Mining Techniques: For Marketing, Sales, and Customer Support*. New York: Wiley, 2011

Berson, A., Smith S., & Thearling K., " Building Data Mining Applications for CRM", NewYork:McGraw-Hill, 2000

Che, Z., Tzu, A., "Feed- Forward Neural Networks Training: A Comparison between genetic algorithm and back- propagation learning algorithm", *Computing, Information and Control ICIC International*, Volume 7, Number 10, October 2011

Diabetes , *How Type2 Diabetes Can Affect Your Body?*. Available at http://diabetesforlifememphis.org/pdf/HowType2DiabetesCanAffectYourBody.pdf

Dunham, H. M., & Linoff, G., *Data Mining: Introductory and Advanced Topics,* New York: Wiley, 2004

Farmer, L., Safer, A., & Chuk, E., "Data Mining Technology across Academic Disciplines*", Intelligent Information Management*, Vol. 3 No. 2, pp. 43-48, 2011

Fayyad, U., Piatetsky G., & Smyth, P. From Data Mining to Knowledge Discovery in Databases AI, 6, *American Association for Artificial Intelligence,* Volume 17 Number 3, 1996

Ferri,C., Orallo, H., & Modroiu, R., *An Experimental Comparison of Performance Measures for Classification* Refaeilzadeh, P., Tang, L., Huan, L., Cross-Validation, Arizona University, 2008

Han, J. & Micheline, K., *Data Mining: Concepts and Techniques: Concepts and techniques*

Hemalatha, M., Megala, S., "Mınıng Techniques in health care: a survey of immunization", *Journal of Theoretical and Applied Information Technology*. Vol:25, No:2, 2011

Hornikova, A., Durakbasa, N. M., Güclü, E. & Bas, G., "Data Mining- Novel Statistical Method", Talk: 8th International Conference on Measurement, Smolenice, Slovakia ,ISBN: 978-80-969-672-4-7; 12 – 15, March 2011

IDF Diabetes Atlas, Fifth Edition

International Diabetes Federations, IDF Diabetes, Six Edition

Jibon, K. "Data Mining: A conceptual Overview", Vol.58, March 2011, pp.7-16, *Communications of the Association for Information Systems*,Volume 8, 2002

Jibon, K., "Usefulness and applications of data mining in extracting information from different perspectives", *Annals of Library and Information Studies*, Vol.58, pp.7-16, March 2011

Kilany, R., (2013). *Efficient Classification and Prediction Algorithms for Biomedical Information*, Doctoral Dissertation, University of Connecticut Graduate School, 2013

Maimon, O. & Rokach, L. *Introduction to knowledge discovery in databases*, 2005

Mao, J., Mahiuddin, K. M. & Jain, K. *Artificial Neural Networks: A Tutorial*, Michigan State University.

Ms. Ishtake S. H. & Prof. Sanap S, A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", *International J. of Healthcare & Biomedical Research*, Volume: 1, Issue: 3, syf. 94-101, April 2013

Parthiban, G. & Srivatsa, S. K., "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients", *Foundation of Computer Science FCS*, New York, USA ,Volume 3– No.7, August 2012

Purnami, S. W., Keputih S.& Zain, J. M.,"Data Mining Technique for Medical Diagnosis Using a New Smooth Support Vector Machine", Zavoral et al (Eds): NDT 2010, Part II, CCIS 88, syf. 15- 27, 2010

Sicree, R., Shaw, J. & Zimmet, P. ,"Diabetes and Impaired Glucose Tolerance", *IDF Diabetes Atlas fourth edition*, 2011

Tabish, S. A., "Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century?", *Int J Health Sci (Qassim)*,1(2): V–VIII, Jul 2007

Tabish, S. A., "National Kidney Disease", *Int J Health Sci (Qassim)*, v.1(2), 2014

Wikipedia., *Weka(Machine Learning)* Available at http://en.wikipedia.org/wiki/Weka_(machine_learning)

Zhang, Y., Hongwen, Z., "Knowledge discovery in astronomical data", *Proc. of SPIE*, Vol. 7019 701938-6, 2011