Adama Science and Technology University

School of Graduate Studies

Department of Computing

INTEGRATED DATA MINING AND KNOWLEDGE BASED SYSTEM FOR PREDICTION AND TREATMENT ADVICE OF DIABETES

Tesfaye Fufa

September 2015

Adama Science and Technology University

School of Graduate Studies

Department of Computing

A Thesis Submitted to the Department of Computing of Adama Science and Technology University in Partial Fulfillment of the Requirements for Degree of Master of Science in Software Engineering

By

Tesfaye Fufa

September 2015

Adama, Ethiopia

Adama Science and Technology University

School of Graduate Studies

Department of Computing

INTEGRATED DATA MINING AND KNOWLEDGE BASED SYSTEM FOR PREDICTION AND TREATMENT ADVICE OF DIABETES

By

Tesfaye Fufa

Names and Signature of Members of the Examining Board

Chair Person, Examining Board	Signature
Dereje Teferi (PhD) Advisor	Signature
Examiner	Signature

ACKNOWLEDGEMENT

First and foremost I would like to thank the almighty God for giving me the courage and strength to finish this study.

Next my best gratitude goes to my advisor Dereje Teferi (PhD) for his, unreserved comments, encouragement, guidance and motivation he gave me to accomplish this thesis.

I sincerely thank my whole family, without their support and encouragement I would have not success today. They have always been with me supporting, helping and appreciating in my journey to be a better man.

In addition, my gratitude goes to all Black Lion Hospital Employees at record offices and Medical Director, Sister Genet record office manager, Dr. Tedla Tsega Diabetes Clinical Center manager and ICT center those helping me to get datasets of diabetes patients.

Last but not least I would like to thank Dr. Temirat Mulugeta from Adama Referal Hospital for his unforgettable help about my thesis by giving suggestion and comments, my best friend Mr. .Garamu Tilahun for his unforgettable encourage in any situation as my family and Mr. Wazih Ahmad from computing department.

Tables of Contents

Lists of Figures

Lists of Listings

Lists of Tables

Lists of Acronyms and Abbreviation

AI: Artificial Intelligence

ARFF: Attribute Relation File Format

ARM: Association Rules Mining

CSV: Common Separated Value

DPAA-W-KBS: Diabetes Prediction and Advisement With knowledge base system

FBG: Fasting Blood Glucose

FBS: Fasting Blood Sugar

IDF: International Diabetes Federation

IDDM: Insulin Dependent Diabetes Mellitus

JPL: Java interface to Prolog

KB: knowledge Base

KBS: Knowledge Base System

KDD: knowledge discovery in database

NIDDM: Non-Insulin Dependent Diabetes Mellitus

RBR: **Rule Based Reasoning**

RQ1: Research Question One

RQ2: Research Question Two

Type1: type one diabetic

Type2: type two diabetic

UAT: User Acceptance Testing

Abstract

Diabetes mellitus is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action or both. It is divided into two major type's type1 and type2 diabetes. During the past decade, diabetes mellitus has emerged as an important clinical and public health problem throughout the world, especially in developing country like Ethiopia (800,000 patients in 2000 E.C) .Extracting implicit knowledge from domain expert about diabetes was so difficult in order to get knowledge about diabetes.

The main objective of this study is to develop a predictive model and integrate the model with Knowledge-Based diabetes prediction system that provides Advisement and decision support for the users.

To build the predictive model dataset is pre-processed for missing values, outliers, noise and errors, based on the 12,139 sample data gathered from Black Lion Diabetes Clinic Center. In this study the model is experimented using J48 decision tree, Jrip, PART, LMT and REPTree. For the knowledge-based part, knowledge is acquired from experts and documents, the knowledge-based system integrating with data mining was developed using WEKA, java, swi-prolog tools.

In this study as compared to other algorithm, the performance of J48 decision tree reveals that 93.84% correct results are achieved for developing classification rules that can be used for prediction and the system achieved 91.6% of the user acceptance. Thus, knowledge discovered with this algorithm is used to integrate with knowledge-based system using Java. As a result, the knowledge-based system is used to predict based on the attribute values.

The researcher concludes that diabetes prediction and advisement should be implemented at any time to be effective it should be supported by data mining technology to overcome on problem of diabetes. However, further investigation should be required for the future to significantly support this struggle.

Keywords: *Data Mining, Knowledge base, Diabetes*

CHAPTER ONE
INTRODUCTION

1. Background

Diabetes mellitus is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action or both (American diabetes association, 2005), manifested by carbohydrates, fat, protein metabolism abnormality. It is a chronic disease, which has no cure yet and associated with high rate of morbidity and mortality in both developing and developed countries and also becoming a pandemic in the world, with increased need for healthcare. Diabetes mellitus is increasingly prevalent and one of the top public health concerns all over the world (Kalayou Kidanu, 2013).

There are two types of well-known diabetes type there are Type1and Type2 diabetes, type1 diabetes, which used to be called juvenile diabetes, develops most often in young people; however, type 1 diabetes can also develop in adults. In type 1 diabetes, the body no longer makes insulin or enough insulin because the body's immune system, which normally protects from infection by getting rid of bacteria, viruses, and other harmful substances, has attacked and destroyed the cells that make insulin, Type 2 diabetes, which used to be called adult-onset diabetes, can affect people at any age, even children. However, type 2 diabetes develops most often in middle-aged and older people. People who are overweight and inactive are also more likely to develop type 2 diabetes. Type 2 diabetes usually begins with insulin resistance a condition that occurs when fat, muscle, and liver cells do not use insulin to carry glucose into the body's cells to use for energy. As a result, the body needs more insulin to help glucose enter the cell (Bethesda, 2013.).

During the past decade, diabetes mellitus has emerged as a serious clinical and public health problem throughout the world .It is estimated that more than 180 million people worldwide have diabetes. It is projected that this number would double by 2030. In 2005, an estimated 1.1 million people died from diabetes and nearly 80% of diabetes deaths occurred in low and middle income countries. Diabetes deaths are projected to increase by over 80% in upper-middle income countries between 2006 and 2015 (Bethesda, 2013.).

According to African at Glance report In Africa, 76% of deaths due to diabetes were in people under the age of 60. IDFA reported Ethiopia to be ranked third among the ten top countries in Africa with 1.4 million Diabetic mellitus cases and estimated prevalence of 3.32% by year 2012 (Port Louis, Mauritius., 2009).In addition Ethiopia is the third ranked with 1.8 million from Top 5 countries for number of people with diabetes (20-79 years), 2013 (Megerssa et al., 2013).

The estimated the number of cases of diabetics in Ethiopia to be about 800,000 in 2000, and Care for diabetic patients may require close and sustained support from a health care team, adequate financial resources, and advanced patient knowledge and motivation. In this respect, there is lack of information in the country. There are more people with diabetes living in urban (246 million) than in rural (136 million) areas although the numbers for rural areas are on the increase. In low- and middle-income Countries, the number of people with diabetes in urban areas is 181 million, while 122 million live in rural areas. By 2035, the difference is expected to widen, with 347 million people living in urban areas and 145 million in rural areas (Glance, Africa at, 2013.). As stated in (ATLAS, 2013).In Ethiopia, national data on prevalence and incidence of diabetes are lacking. However, patient attendance rates and medical admissions in major hospitals are rising. The estimated prevalence of diabetes mellitus in adult population of Ethiopia is 1.9% (4, 5). Moreover, Cohen et al reported a high prevalence of diabetes (8.9%) among young (age < 30 years) in Ethiopian Jews who have been to Israel for less than 4 years .WHO estimated the number of diabetic cases in Ethiopia to be 800,000 by the year 2000, and the number is expected to increase to 1.8 million by 2030 (Getachew, 2014.).

The disease condition requires competent self-care, which can be developed from a thorough understanding of the disease process and the management challenges by the patient and family members. As a result of this, diabetic education and counseling for the patient and family members are becoming important goals of diabetic patients care today. Unfortunately, about a third of the people suffering from diabetes may not be aware of it early considering the insidious onset and development (Naheed G., 2010).

1.1. Knowledge based system

Knowledge-Based System (KBS) is one of the major family members of the AI (artificial intelligence) group. With availability of advanced computing facilities and other resources, attention is now turning to more and more demanding tasks, which might require intelligence. The society and industry are becoming knowledge oriented and rely on different experts' decision-making ability. KBS can act as an expert on demand without wasting time, anytime and anywhere. KBS can save money by leveraging expert, allowing users to function at higher level and promoting consistency. One may consider the KBS as productive tool, having knowledge of more than one expert for long period of time. In fact, a KBS is a computer based system, which uses and generates knowledge from data, information and knowledge. These systems are capable of understanding the information under process and can suggest decision based on the residing knowledge in the system whereas the traditional computer systems do not know or understand the information they process.

Knowledge-based systems are more useful in many situations than the traditional computer based information systems. Some major situations include: when expert is not available, when expertise is to be stored for future use, When intelligent assistance and training are required for the decision making for problem solving, When more than one experts' knowledge have to be combined at one platform (Akerkar, Eds. Sajja, 2010).However, the scarcity and nature of knowledge make the KBS development process difficult and complex. The transparent and abstract nature of knowledge is mainly responsible for this. In addition, this field needs more guidelines to accelerate the development process. There are some of the major limitations with the KBS: Acquisition, representation and manipulation of the large volume of the data/information/ knowledge, High-tech image of the AI field, Abstract nature of the knowledge, Limitations of cognitive science and other scientific methods (Akerkar, Eds. Sajja, 2010).

1.2. Data Mining

Data mining is the process of analyzing data from different perspectives and extract unknown information it into useful information, information that can be used for industrial, medical and scientific purposes. As such the process of data mining involves sorting through large amounts of data and discovering patterns in the data (Velide, 2014).

Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis and medical devices etc. The large amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Therefore Data mining brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions (Tan, 2005).

1.3. Statement of the Problem

In Ethiopia number of professional existing at health care is less and number of patients is rising every year at all regions. According to IDFA reported Ethiopia to be ranked third among the ten top countries in Africa with 1.4 million diabetes cases and estimated prevalence of 3.32% by year 2012. In addition Ethiopia is the third ranked with 1.8 million from Top 5 countries for number of people with diabetes (20-79 years). Most of the Ethiopian people (85%) are farmers and live in villages which are often remote and inaccessible. Surfaced roads link the main centers, but many places can be reached only by footpaths. Transport is either by bus and taxi, which is relatively expensive, or by mule and foot, which is laborious and slow. Indeed, the delivery of expertise and medicines is a complex and largely unresolved problem, which is magnified in the delivery of Advisement for chronic lifelong diseases (which are increasingly being recognized) such as diabetes (al Brown, 2013).

As (Solomon Gebremariam, 2013) suggested that tacit knowledge about the diagnosis and Advisement of diabetes is extracted from the domain experts using interviewing method in order to have detail understanding of the domain knowledge. However, it was a difficult task to extract the necessary knowledge due to the personal nature of tacit knowledge.

As described by (A. Ranjan, 2002) , acquiring knowledge from experts or domain knowledge is not an easy task. The following are some factors that add to the Complexity of knowledge acquisition from experts and its transfer to a computer:

- Experts may not know how to articulate their knowledge or may be unable to do so.
- Experts may lack time or may be unwilling to cooperate.
- Testing and refining knowledge is complicated.
- Methods for knowledge elicitation may be poorly defined.
- System builders tend to collect knowledge from one source, but the relevant knowledge may be scattered across several sources.
- Builders may attempt to collect documented knowledge rather than use experts.
- The knowledge collected may be incomplete.
- It may be difficult to recognize specific knowledge when it is mixed up with irrelevant data.
- Experts may change their behavior when they are observed or interviewed.
- Problematic interpersonal communication factors may affect the knowledge engineer and the expert.

A critical problem in the development of knowledge-based systems is capturing knowledge from the experts. There are many knowledge elicitation techniques that might aid this process, but the fundamental problem remains: tacit knowledge that is normally implicit, inside the expert's head, must be externalized and made explicit. Knowledge acquisition (KA) thus has been well recognized as a bottleneck in the development of knowledge based systems and is a key issue in knowledge engineering. Traditionally, KA techniques can be grouped into three categories: manual, semi-automated (interactive) and automated (machine learning (ML) and data mining). Since the early days of artificial intelligence (AI), the problem of KA, the elicitation of expert knowledge in building knowledge bases, has been recognized as a fundamental issue in knowledge engineering (Tu Bao Ho, 2007).

Therefore the following questions should be answered at the end of this study:

RQ1: Which data mining techniques is the best for predicting of diabetes types?

RQ2: How to integrate Data Mining and Knowledge base?

1.4. Objective of the Study

The general and specific objectives of this study are described here under.

1.4.1 General objective

The general objective of this study is to explore the possibility of integrating data mining techniques with knowledge bases system for diagnosis and treatment of diabetes patients.

1.4.2 Specific objectives

The following specific objectives are incorporated for achieving the general objective.

- To review literatures on the concepts of integrating knowledge based system with data mining techniques and about diabetes.
- To apply data mining techniques in order to extract hidden information to get knowledge for decision making.
- To acquire knowledge about major diabetes diseases, symptoms and diagnosis from primary and secondary sources using interview, detailed discussion with domain experts and document analysis.
- To develop prototype that integrates data mining techniques with knowledge base system to diagnose or predict and Advisement for diabetes.
- To test and evaluate the performance and user acceptance of the developed prototype.

1.5. Significance of the Study

Identifying problems and attempting to offer solutions is a significant step in any sort of investigation. The study will create awareness of making effective communication between the expertise and the patient's .Using the recommended solutions, hospitals and clinic at remote areas will also be initiated to use the designed prototype of the integration of data mining techniques and knowledge base system. The findings of this study will benefit health care experts in such a way that experts and patients can easily get disease identification and diagnosis system to manage diabetes diseases from the knowledge base which stores the facts and rules that experts used to solve problem.

Knowledge-based systems can be used to enhance problem solving capacity in a flexible manner. Such systems can also document knowledge for future use and training. This leads to have increased quality in problem solving process. It is also advantageous for remote and rural areas that have scarcity of medical experts and solves the problems of knowledge acquisition for knowledge engineer. In general, the benefits and beneficiaries of the research will be the following:

Benefits of the research:

- The system will help diabetes diseases patients and physicians by providing useful information about the disease identification, diagnosis and Advisement.
- The system will help non experts in diabetes diseases identification and diagnosis work. Non experts can improve their knowledge and be effective in diabetes diseases diagnosis and Advisement as expert knowledge is readily available.

The beneficiaries of the proposed system will be the following:

- Patients , individuals, the public in general or community
- Hospital professionals /experts
- Governmental and non-governmental institutions, Health institutions, etc.

1.6. Scope and limitations

The scope of the study is limited to integrating data mining and knowledge based system in order to diagnose and advisement diabetes disease. It also helps to understand type1 and type2 disease and their advisement mechanisms to facilitate the awareness for patients and physicians. The study did not include other types of diabetes like gestational diabetes. The researcher interview with domain experts revealed that this diabetes type is seen especially in pregnant women's and it is related with type two diabetes complication. If she has type two diabetic before pregnancy, there may be the probability to develop gestational diabetes type. Therefore the study cannot include these types of diabetic and the researcher only uses the datasets from one center because the other hospital or clinics have no separated data about diabetes. The system only gives the advice for the users; it cannot gives treatment for the patient like experts, for instance injection when the injection is needed.

1.7. Methodology of the Study

Methodology provides an understanding of how a research is conducted and organized in order to obtain information that is helpful for developing the design of a research. The methods used in this research are described below.

1.7.1. Interviewing Domain Experts

Both organized and non-formal interviews were employed to elicit tacit knowledge from domain experts. Since one of the specific objectives of this research is extracting tacit knowledge which is entrenched and personalized in experts mind. For this research, four experts were selected purposively for interview about diabetes.

1.7.2. Literature Review

In order to have deep understanding of the problem it is vital to review several literatures that have been conducted in the field so far. For this reason, related literature such as books, articles, proceeding papers and some other sources that are retrieved from the Internet have been

consulted so as to understand the domain knowledge, concepts, principles and methods that are important for developing knowledge-based systems.

1.7.3. Knowledge discovery in database (KDD) process

Data mining is often set in the broader context of KDD. The KDD process basically involves selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures (David Hand, 2001). Sample data was taken from Black lion Diabetes Clinic Center with data types of integers and nominal values with 12,139 sample data for this study.

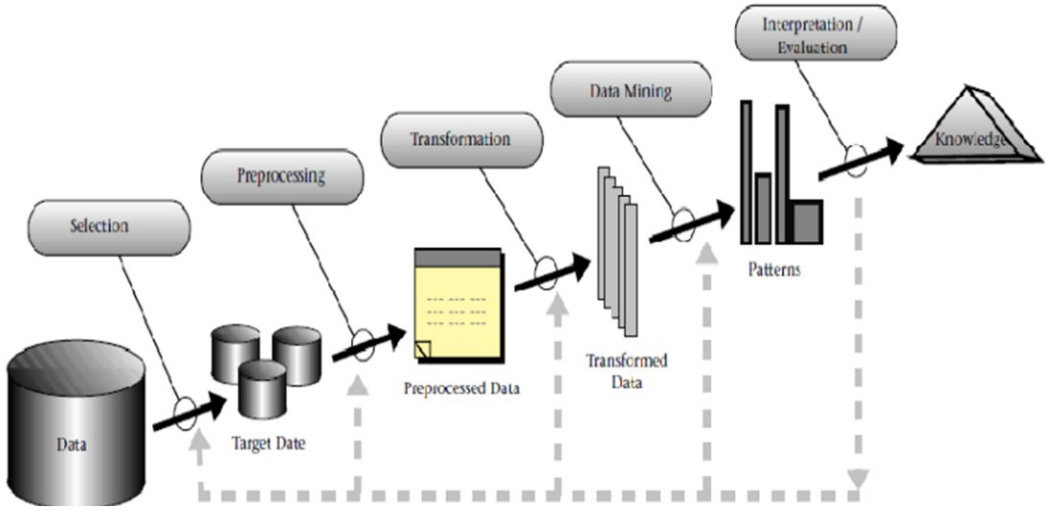

Figure 1-1 Knowledge discovery in database (KDD) process, adapted from (David Hand, 2001)

As figure 1.1 shows several steps from selecting data to providing understandable knowledge to the user are involved in the KDD process. These are further explained in the following steps.

Step 1. **Selection**: Before directly going to selection one needs to understand the domain and identify the goal of applying the process. Then select the necessary data to solve the problem. The main part of the KDD process is the selection of raw data that is necessary for the discovery. There might be unnecessary data attributes provided, but only few data attributes are needed in the process. Selecting the necessary data attributes for the discovery places an important part which yields target dataset. The diabetes data was selected based on the attribute importance to predict diabetes types after discussing with domain experts.

Step 2. **Preprocessing**: Handling missing values, eliminating noise and duplicate records in the data sets is the main part of this process. Outlier detection and Missing values in the data lead to loss of useful information, which might not result in discovering useful knowledge. Noise in data and duplicate records mislead the process in obtaining accurate knowledge. Therefore, data cleaning and the preprocessing are necessary to produce better quality results and the researcher applied filling missing values from filters preprocessing panel. There are various tools used for this study such as Microsoft Excel and WEKA 3.6.9 that can be applied for these tasks.

Step 3. **Transformation**: In this step data is transformed or consolidated into forms appropriate for data mining. Data transformation involves the following. First, smoothing, where the noise from data is removed, second, aggregation, where aggregation operations are applied to data for the analysis of the data, third, generalization, where raw data is replaced with high level concepts. Lastly, normalization, where the attribute data are scaled to fall within a specified range; and feature Selection, where new attributes are constructed and added from the given set of attributes. Therefore for this study from different attributes age and fasting blood sugar was transformed into nominal data types based on the ranges given as the experts are using to diagnose diabetes types.

Step 4. Data mining: The step that requires analysis of the main problem and decision on which models and parameters are appropriate. Depending on the model, different data mining algorithms and methods are chosen that are needed for searching data patterns. Data mining methods are performed to achieve the goal by finding the interesting patterns in the data. Better results are obtained if the preceding steps are performed properly. Therefore for this study classifier algorithm is applied for data mining purpose and due to the performance score, j48 classifier was used.

Step 5. Interpretation/Evaluation: The step where the mining results are interpreted and even the process can start again from step 1 if there are any errors or for providing further accurate results. It involves the visualization of extracted patterns and results. The knowledge discovery process then takes the raw results from data mining and transforms them into useful and understandable information for the users. Using the knowledge directly, incorporating the

knowledge into another system for further action, or simply documenting it and reporting it to interested parties, checking for and resolving potential conflicts with previously believed or extracted knowledge can be part of this step. Five classifier algorithm like J48, Jrip, PART, LMT and REPTree was applied on the dataset and each algorithm are evaluated based on different criterion like: Precision , Recall, True Positive ,Correctly classified and incorrectly classified instances etc.

1.7.4. Evaluation methods

The set of exposed rules has to be verified for accuracy (the rules reveal the dataset), consistency (no redundant or contradictory rules) and usefulness (rules showing the decision making process) for the knowledge base to be developed (O, M. Mehdi Owrang, 2008). The accuracy of the models developed using data mining techniques are evaluated based on prediction accuracy of classifiers, Precision, Recall, F-measure and True Positive rate. Moreover, it is also tested by users to ensure user acceptance and check the extent to which the system meets user requirements.

1.8. User Acceptance Testing (UAT)

User acceptance testing (UAT) is the last phase of the software testing process**,** the aim of undertaking user acceptance testing is to make sure how well DPAT-KBS is performing on the eyes of users so as to make sure that the system is accepted and usable by users. Five domain experts were selected to test the system.

1.9. Knowledge Representation

There are many techniques used for knowledge representation. In this research, rule based method was used to represent expert knowledge concerning diabetes diseases diagnose and Advisement techniques. In a rule based system much of the knowledge is represented as a rule that is as conditional sentences relating statements of facts with one another. Rules are constructed in the form of if-then format from the j48 tree generated rules.

1.9.1. System Development and Implementation Tools

Weka tool is used for the data mining in order to extract hidden information for knowledge discovering from large data, Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning research written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality (Dr. Sudhir B. Jagtap, 2013).

Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from one's own Java code (Dr. Sudhir B. Jagtap, 2013).

The main reason weka is used for this research are as follows:

- It is freely available under the GNU General Public License
- It is Portable, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- It contains collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces

Prolog programming language is used for knowledge representation because this programming language is used to develop prototype knowledge based system that diagnoses diabetes diseases. Prolog (programming in logic) is one of the most widely used programming language, especially in the artificial intelligence research. Prolog is a declarative language and has the capacity to describe the real world. In order to integrate data mining with knowledge based system java Neatbeanse (IDE 8.2) with JDK 8.

1.10. Organization of the thesis

This thesis includes of seven chapters. **Chapter one** discusses an introduction about the study giving highlight about the area to which the study is focused. It provides highlight about diabetes, data mining and knowledge based system. Problem statement, objective, significance, scope, and methodology used in the study are also included.

Chapter two is basically dedicated for literature review. In this chapter, a detailed discussion about data mining and its tasks pertinent for this study are included. The concept of diabetes, types and the advantages of data mining techniques are discussed. Moreover, since the concern of the study is an integration of data mining and knowledge base system for diabetes prediction and Advisement, literatures focused on employing data mining model for construction of knowledge based system are discussed. Discussion about knowledge base system including types of reasoning for knowledge base system, how knowledge is acquired and represented so as to develop knowledge base system is also covered. In addition, related works in the area of diabetes by using data mining techniques and knowledge base are also included in this chapter.

Chapter three presents the knowledge acquisition process using data mining techniques. The focus here is on automatic knowledge acquisition techniques through data mining. Knowledge discovery steps such as data set preparation, preprocessing and predictive model creation as well as experimentation are also discussed in the chapter. The results of WEKA classifier algorithms are analyzed, interpreted, evaluated and compared with each other and the best model is selected for this study.

Chapter four discusses Knowledge acquisition, modeling and representation about the diabetes Advisement and managements by interviewing the domain experts.

Chapter five is all about the integration of data mining result with knowledge base system architecture with detail discussion.

Chapter six discuss about implementation, experimentation and the basic functionality of the knowledge base, performance evaluation and user acceptance testing are discussed under this chapter.

Finally, **Chapter seven** is dedicated for conclusion and recommendation. In this chapter based on the result obtained from the study the researcher‘s concluding remarks and recommendation for future works are presented.

CHAPTER TWO

LITERATURE REVIEW

2. Data Mining and Knowledge Based System

2.1. Overview of Data Mining

Data mining is the process of examining data from different perspectives and extract unknown information it into Useful information, information that can be used for industrial, medical and scientific purposes. As such the process of data mining involves sorting through large amounts of data and determining patterns in the data (Velide, 2014).

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. The first and simplest analytical step in data mining is to describe the data — summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables (such as values that often occur together) (Corporation, 1999).

Data mining is also sometimes called data knowledge discovery, is the process of analyzing data from different perspectives and summarizing it into useful information and that information can be used to increase revenue and reduce costs, or for both at the same time. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different ways, classify it, and summarize the relationships identified. It can be defined as; it is the process of finding correlations or patterns among dozens of fields in large relational databases. These methods allow one to acquire from data, previously hidden knowledge about relations and patterns in behavior of the investigated object (Kohavi, 2001).

Past decades has seen a dramatic increase in the amount of information of data being stored in different electronic format. The problem is how to maintain these data and also extraction of required knowledge. Data mining techniques can be viewed as a result of the natural evaluation of information technology, the process of analyzing the observational data sets to find unsuspected relationships and to summarize the data in novel ways those are both understandable and useful to the data user called as data mining (R. Jayabrabu, 2012).

2.1.1. Objective of data mining

As (Jackson, 1999) explain, the objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing , Finding useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing) .The term “data mining” is primarily used by statisticians, database researchers, and the MIS and business communities. The term Knowledge Discovery in Databases (KDD) is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process The additional steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results of the data mining process, ensure that useful knowledge is derived from the data.

The ultimate objective of data mining is knowledge discovery. Data mining methodology extracts hidden information from large databases. With such a broad definition, however, an online analytical processing (OLAP) product or a statistical package could qualify as a data mining tool (Gilman, 2000).

2.1.2. Data mining application in health care

Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. The large amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Therefore, Data mining brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions (Tan, 2005).

According to (Tan, 2005) the advantages of data mining in healthcare include:

- Healthcare insurers detect fraud and abuse,
- Healthcare organizations make customer relationship management decisions,
- Physicians identify effective Advisement and best practices, and
- Patients receive better and more affordable healthcare services.

As described by (M. Durairaj, V. Ranjani, 2013) the application of data mining in health care are as follows:

Advisement effectiveness

Data mining applications can develop to evaluate the effectiveness of medical Advisement. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of Advisement. Advisement plan (to patients) Data mining could be particularly useful in medicine when there is no dispositive evidence favoring a particular Advisement option, Based on patients' profile, history, physical examination, diagnosis and utilizing previous Advisement patterns, new Advisement plans can be effectively suggested by data mining methods.

Healthcare management

Data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of Hospital admissions and claims to aid healthcare management. Data mining used to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists.

Customer relationship management

Customer relationship management is a core approach to managing interactions between commercial organizations typically banks and retailers-and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings.

Fraud and abuse

Detect fraud and abuses establish norms and then identify unusual or abnormal patterns of claims by physicians, clinics, or others attempt in data mining applications. Data mining applications fraud and abuse applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims.

Medical Device Industry

Healthcare system's one important point is medical device. For best communication work this one is mostly used. Mobile communications and low-cost of wireless biosensors have paved the way for development of mobile healthcare applications that supply a convenient, safe and constant way of monitoring of vital signs of patients (Raf.R, 2012).

Hospital Management

Organizations including modern hospitals are capable of generating and collecting a huge amount of data. Application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized (ShusakuTsumoto and Shoji Hirano, 2011).

2.1.3. Data Mining algorithms

Data mining algorithms are used to specify the kind of patterns to be found in data. Data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database (Han, J & Kamber, M, 2006)the goal of descriptive data mining is to gain an understanding of the analyzed system by discovering patterns and relationships in large datasets (Kantardzic, 2003). Predictive mining tasks, on the other hand, perform inference on the current data in order to make predictions (Han, J & Kamber, M, 2006). Classification is the commonly used data mining technique for prediction, whereas association and clustering are considered the descriptive data mining techniques (Kantardzic, 2003). Therefore the followings are data mining classifier algorithms.

Decision Tree

Decision tree models have been around for a very long time, although formal methods of building them are a relatively recent innovation. Before the development of such methods they were constructed on the basis of prior human understanding of the underlying processes and phenomena generating the data (Hand, 2001).

Decision trees are supervised learning methods that construct decision trees from a set of input-output samples. A typical decision tree learning system adopts a top-down strategy that searches for a solution in a part of the search space (Kantardzic, 2003). Decision tree induction is the learning of decision trees from class-labeled training dataset (Han, J & Kamber, M, 2006). There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literature (Kantardzic, 2003). ID3, C4.5, J48 and Classification and Regression Trees (CART) are the algorithms used to construct decision trees in a top-down recursive divide-and-conquer manner (Han, J & Kamber, M, 2006). As stated by (Kantardzic, 2003), a well-known tree-growing algorithm for generating decision trees is Quinlan's ID3 with an extended version called C4.5.

According to (Larose, Daniel T, 2005) there are requirements to be met before implementation of decision tree algorithms. First, since decision tree algorithms represent supervised learning they require predetermined target variables and a training dataset which provides the algorithm with the values of the target variable. Second, the training dataset should be rich and varied, providing the algorithm with a healthy cross section of the types of records for which classification may be needed in the future.

Decision trees learn by example, and if training set are systematically lacking for a definable subset of records, classification and prediction for this subset will be problematic or impossible. Third, the target attribute classes must be discrete values that are clearly demarcated as either belonging to a particular class or not belonging. Decision tree classifiers have good accuracy, in general. However, successful use may depend on the data at hand. During tree construction, attribute selection measures are used to select the attribute that best partitions the tuples into distinct classes (Han, J & Kamber, M, 2006).

Rule-Based Classification

Even though the pruned trees are more compact than the originals, they can still be very complex. Hence, to make a decision tree model more readable, it can be transformed into an IF-THEN decision rule (Kantardzic, 2003). As stated by (Larose, Daniel T, 2005). Decision rules can be generated from a decision tree by traversing any given path from the root node to any leaf. The complete set of decision rules generated by a decision tree is equivalent to the decision tree itself. A rule-based or decision rule classifier such as C4.5 uses a set of IF-THEN rules for classification. An **IF-THEN rule** is an expression of the form IF condition THEN conclusion. If the condition in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied and that the rule covers the tuples (Han, J & Kamber, M, 2006).

2.2. Knowledge based system

Knowledge-Based System (KBS) is one of the major family members of the AI group. KBS can act as an expert on demand without wasting time, anytime and anywhere. KBS can save money by leveraging human expert, allowing users to function at higher level and promote consistency of work. One may consider the KBS as productive tool that have knowledge of more than on expert for long period of time. In fact a KBS is a computer based system which uses and generates knowledge from domain expert (Akerkar, P. Heijst., 2010).

Knowledge-based system (KBS) is a computer program that is derived from a computer science field of study known as Artificial Intelligence (AI). The systematic aim of AI is understanding intelligence by developing computer programs that show intelligent behavior. It deals with methods of inferring mechanisms using the computer and in what way knowledge can be represented using different techniques for inferring (E.A. Feigenbaum and R.S. Engelmore, 1993).

2.2.1. Objectives of Knowledge-Based System

As (E. Turban, 1990)] noted, one example of the specialized branches of AI is KBS. The objectives of KBS are listed as follows:

- Offers a great level of intelligence
- Supports individuals in finding out and constructing new fields
- provides a huge amount of knowledge in several areas
- Assists in knowledge management that is deposited in the knowledge base
- Finds solution for social problems in a well manner than the traditional computer-based information systems.
- Gains new observations by imitating new circumstances
- Provides important software development productiveness
- Importantly decreases time and money to construct automated systems.

2.2.2. Advantages of Knowledge-Based System

The main advantages of using a knowledge-based system are descried as follows (R.A. Akerkar, 2010).

Permanent documentation of knowledge: A knowledge engineer extracts knowledge from domain experts and relevant documents for a certain problem domain and represents it using one of the knowledge representation techniques and transfers it into the knowledge base. This helps end-users to use the knowledge stored for a long-term from the documented knowledge in the knowledge base at any time.

Cheaper solution and easy availability of knowledge: It is assumed very huge, complicated and expensive to develop a KBS. Nevertheless, it costs once for building the knowledge base. After duplicating it into many copies, it is simple to use the knowledge in many places. This interrupts the dominations of domain experts and makes simple to acquire and utilize knowledge. On the contrary, educating new domain experts is inefficient and costly. Therefore, the aim of developing KBSs is to reduce cost, time, human expertise and medical error.

Dual advantages of effectiveness and efficiency: Since knowledge-based systems are computer-based systems, they have efficiency-directed factors such as speed, accuracy, control, and permanent content storage. It is possible to make the knowledge-based system effective by integrating the knowledge element. They are more efficient than domain experts and attempts to become equally effective like domain experts.

Consistency and reliability: Since the knowledge element is integrated into the KBS and the capability to perform effectively, the trustworthiness of the system rises. Besides, dupery and errors can be stopped. Information can be accessible rapidly for making decision with appropriate justification. As the level and amount of knowledge rises, making right decision will rise and thereby reduce the threat of wrong decision.

Justification for better understanding: The reliability of the domain experts relies on the capability to explain their decisions. This can be offered by using the reasoning and justification component of the KBS to the end-users. If there is a well understanding of the decisions made by end-users of the system, then it increases the quality and trustworthiness of the system.

Self-learning and ease of updates: With the assistance of the inference engine of the system, the knowledge base always updates its knowledge from experience. The knowledge-based system can update its knowledge either using automatic machine learning or manually by the knowledge engineer. Such self-learning advances the adaptability and tractability of the system.

2.2.3. Limitations of Knowledge-Based System

The following are some of the major drawbacks of using the KBS as described by (R.A. Akerkar, 2010).

Partial self-learning: The knowledge-based systems can explain when they made a decision and learn from experience by updating its knowledge. However the represented knowledge may not be completely known so that the knowledge-based system can learn partially from experience. Besides, domain experts can conform automatically to new conditions though KBSs should explicitly update their knowledge.

Creativity and innovation: It is not possible computer-machines to show a certain behavior as creative as humans do. Domain experts can answer back in a creative manner to new conditions though knowledge-based systems as a maximum can deal with the five basic senses. KBSs do not have any methodology to deal with invention, the ability to create and common sense. If we use AI methods in KBS, humanlike five basic senses can be partly applied. The vision, listening, smell, taste, and touch tasks are implemented so that they cannot totally assist activities associated to perception, emotion, and enjoyment. This is because knowledge-based systems are now reliant on symbolic input though human beings have a varied of sensory experience.

Weak support of methods and heuristics: Knowledge-based systems cannot operate with their full capacity if there is no response given or the problem is out of the system's knowledge. When the heuristics is applied to look for a solution from the search space, the success of the systems relies on the quality of the heuristics. Thus, the responsibility depends on the knowledge engineer to develop the heuristics.

Development methodology: System development is not only an art but also a science. For example, in the development of information systems there is no one common accepted methodology. There are common guidelines and lifecycle models that help to develop all types of computer-based information systems. However, there is no common development model that helps knowledge engineers to develop a KBS.

Knowledge acquisition: It is the transfer of knowledge from its source into an appropriate format that can be used by the knowledge-based system. Knowledge is basically personal in nature and is therefore very challenging to extract the embedded knowledge from human mind.

Development of testing and certifying strategies and standards for knowledge-based systems: A knowledge-based system operates in a specific problem domain. The absence of standardization is the main limitation of the existing state of acceptance of knowledge-based system. The acquired knowledge from its source should be tested before representing it into the knowledge base using one of the knowledge representation techniques. Likewise, the knowledge base should be tested even after the representation of the validated knowledge. Standards such as verification, validation and quality metrics are required for ensuring the quality of the KBS.

2.2.4. Architecture of Knowledge based system

Figure 2.1 below shows the building blocks of knowledge based system architecture.

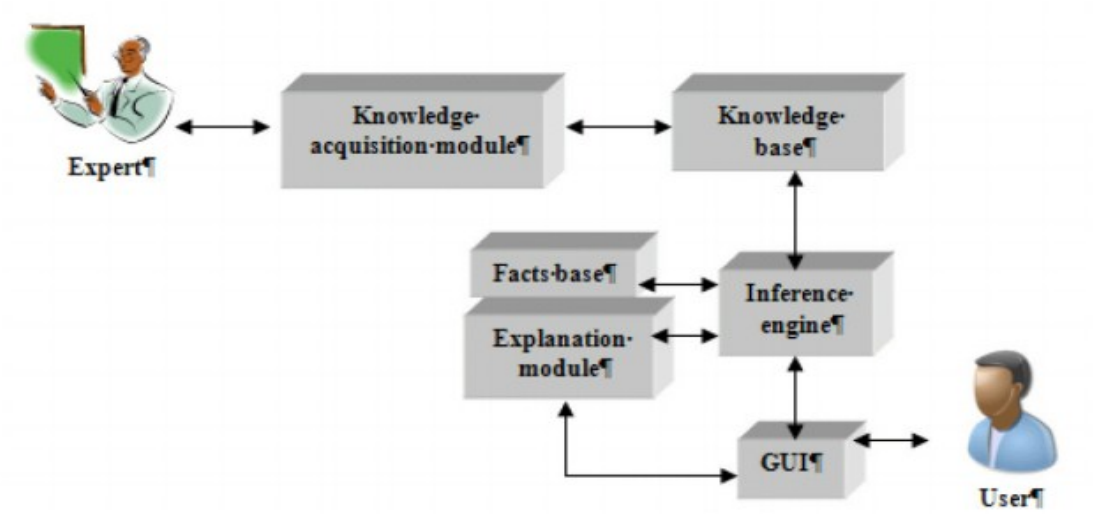

Figure 2-1 Architecture of knowledge based system, adapted from (Saxil, 2010)

The architecture of knowledge based system consists of different components such as Knowledge Base, Knowledge acquisition module, inference engine, user interface and explanation module. The Knowledge Base contains all relevant knowledge acquired from domain experts. Knowledge also acquired from the user during their interaction with the system. The knowledge acquisition module helps in the collection process of knowledge from the set of human experts as shown in Figure 2.1 above. The inference engines formulate questions and assert the answers provided by the user in a natural language form. It provides a mechanism for conveying recommendations to the end user. The explanation module provides a brief description to the user why the system arrived at a certain conclusion (Aref, 2000).

2.2.5. The Knowledge engineering process

The development of knowledge based system is the integration of many components. Figure 2.2 below shows the overview of knowledge based system development process.

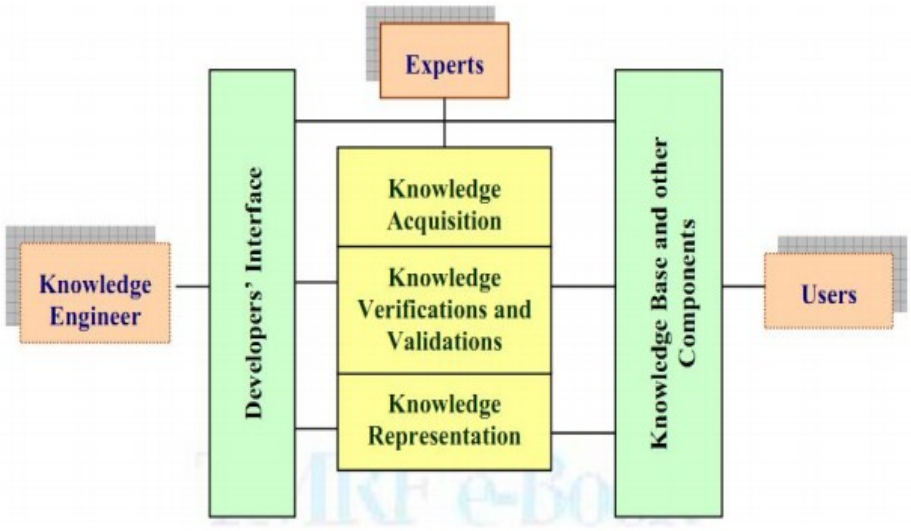

Figure 2-2 Development of a Knowledge-Based System, adapted from (Akerkar P. , 2010)

2.2.5.1. Knowledge acquisition

Knowledge acquisition is the process of acquiring relevant knowledge from human experts, books, documents, sensors, or computer files. The knowledge can be specific to the problem domain or to the problem-solving procedures, it can be general knowledge (e.g., knowledge about business) or it can be meta-knowledge (knowledge about knowledge). Knowledge acquisition is the bottleneck in knowledge based system development today. Because, the trustworthiness and the performance of the knowledge based system mainly depends upon the acquired knowledge (Alechina, 2012).

The knowledge acquisition process incorporates different methods such as interviews, questionnaires, record reviews and observation to acquire factual and explicit knowledge (Gilbar, 1999).

The performance of the expert systems depends upon the reliability, validity and accuracy of the elicited knowledge. The process of knowledge elicitation is affected by different contributing factors such as communication between the expert and ability of knowledge engineer (Tehrani, 2009).

Therefore, effective elicitation techniques facilitate to acquire relevant knowledge form domain experts and by data mining techniques. The commonly used knowledge acquisition techniques are discusses as follows (Wang, 2011) .

a) Interview: An interview technique is the process of interacting with domain expert on how they perform their tasks based on their expertise. Knowledge acquired through direct elicitation methods are procedural knowledge. Based on its structure, interview can be classified into structure, semi structure and unstructured interview (Henok & Burge, 2011, 1998).

Structured Interviews- A structured interview method is questioning the domain expert directly. It is goal-oriented process. It forces organized communication between the knowledge engineer and the domain expert. The structure reduces the interpretation problems inherent in unstructured interviews and allows the knowledge engineer to

Prevent the bias caused by the subjectivity of the domain expert as described by (Ranjan.J, 2006).

A semi-structured interview is an interview which has a guide that usually includes both closed-ended and open-ended questions. It is more flexible than structured one. In these kinds of interview the interviewer has a chance to change the order of questions and expand the dimension of questions based on the participants' responses (Ranjan.J, 2006).

Unstructured Interviews- sessions are conducted informally, usually as a starting point. Unstructured interview techniques provide complete or well-organized descriptions of cognitive processes. There are many reasons that enforced to applying unstructured interview. Domain the experts usually find it very difficult to express some of the most important elements of their knowledge. Through structured interview it is difficult to acquire the required knowledge. With

good training and personal experience knowledge engineers can use unstructured interview to acquire relevant knowledge from domain expert (Ranjan.J, 2006).Therefore, efficient and effective interview techniques largely depend on the ability of knowledge engineer to articulate their implicit knowledge. Because every interview is different in very specific ways and it is difficult to provide comprehensive guidelines for the entire interview process. Therefore, interpersonal communication and analytic skills of knowledge engineer is very important. On the other hand eliciting knowledge using indirect methods requires human intervention such as observation, document analysis (Wang, 2011).

b. Observation

In many ways, this is the most obvious and straightforward approach to knowledge acquisition. Through this technique knowledge engineer directly observe how problems are solved by domain expert. Observation was particularly fruitful in seeing what the expert physically does when solving diagnostic problems. It was useful in determining what types of knowledge the human expert used to solve problems and the forms in which the knowledge were stored. Observations are used primarily as a way of supporting verbal protocols. Generally, acquiring knowledge through observation is expensive and time-consuming task (Gau, 1990).

c. Document Analysis

The final form of knowledge acquisition method is concerned with a detailed analysis of the existing document. This technique is used to collect relevant knowledge from the existed documents of different format. These documents include professional literature, brochures, manuals, guidelines, employee handbooks, reports, glossaries, course texts, and other relevant materials (Gau, 1990).

Knowledge elicitation methods can be classified into different types. Direct and indirect is the commonly known methods of knowledge elicitation. The way of classification depends upon how knowledge engineer directly obtains information from the domain expert (Osuagwu, 2006).

A direct method involves directly questioning a domain expert on how they do their job. In order to implement direct methods successfully, the domain expert has to reasonably articulate and willing to share his/her knowledge. However, in case of indirect methods the required knowledge is not requested directly. Instead, the result of the knowledge elicitation session must

be analyzed in order to extract the required knowledge. Indirect methods are thought to be more suitable when knowledge is not easily expressed by the domain expert (Osuagwu, 2006).

d) Data mining

Knowledge discovery by data mining is a process that extracts implicit, potentially useful or previously unknown information from the data. The ultimate goal of knowledge acquiring by data mining is to find the patterns that are hidden among the huge sets of data and interpret them to useful knowledge and information.

2.2.5.2. Knowledge representation

Acquired knowledge is structured so that it was ready for use in the process of knowledge representation. This activity involves preparation of acknowledge map and encoding of knowledge in the knowledge base.

Knowledge validation: Knowledge validation (or verification) involves validating and verifying the content of knowledge (e.g., by using test cases and confusion matrix) and user acceptance. The testing result of knowledge based system was validated by domain expert.

Inference- This activity involves the design of software to enable the computer to make inferences based on the stored knowledge for the specific domain problem. In other word inference engine is a programs that reason over extensive knowledge bases.

Explanation- This step involves the design and programming of an explanation facility. Explanation module is program that answered how the knowledge based system arrived at certain conclusion. Explanation module addresses the issues of system user interactivity.

2.2.6. Knowledge based reasoning techniques

There are a lots of knowledge based reasoning techniques, some of them are: fuzzy logic, semantic network, neural network, rule based reasoning and case based reasoning techniques. From those techniques case based and rule based are discussed below:

2.2.6.1 Case based Reasoning

Case-based reasoning (CBR) means "adapting old solutions to meet new demands, using old cases to account for new situations, using old cases to evaluate new solutions, or reasoning from precedents to interpret a new situation" (Chen, S., 2007).CBR is more comfortable to make better Decision in dynamically changing environment. People learn from their success and wrong activities to handle similar situations in the right manner and not to repeat their mistake of the past. CBR approach is more compatible to reuse previously solved problems and learning from experiences for future decision (Salem, 2007).Similarly, CBR is an approach to incremental learning. Once a problem has been solved, CBR approaches use the solution to solve for future problems (A. Aamodt, E. Plaza, 1994).

Advantage of case based reasoning

A case based reasoning approach has tremendous advantages in the development of knowledge based system. The following are main the advantages of case-based reasoning (Simon, 2004).

- Ability to express specialized knowledge.
- Naturalness of representation
- Modularity
- Easy to knowledge acquisition
- Self-updatability.
- Handling unexpected or missing values.
- Inference efficiency.

Disadvantage of case based reasoning

Even though case based reasoning approaches have a numbers advantage. But, due to lack of sufficient cases, the construction and inference mechanism of a case-based system loss the required objective. Some of the limitation issues in case-based reasoning are (Prentzas, 2007):

- Inability to express general knowledge
- Knowledge acquisition problems
- Inference efficiency problems and Provision of explanations

2.2.6.2. Rule based reasoning

Rule based reasoning is a system whose knowledge representation in a set of rules and facts. Symbolic rules are one of the most popular knowledge representation and reasoning methods. This popularity is mainly due their naturalness, which facilitates comprehension of the represented knowledge. The basic forms of a rule, **If<condition> then <conclusion>** where <condition> represents premises and <conclusion> represents associated action for the premises. The condition of rules are connected between each other with logical connectives such as, AND, OR, NOT, etc., thus forming a logical function. When sufficient conditions of a rule are satisfied, then the conclusion is derived and the rule is said to be fired.

Rules based reasoning was dominantly applied to represent general knowledge. Rule based expert systems have a significant role in many different domain areas such as medical diagnosis, electronic troubleshooting and data interpretations. A typical rule based system consists of a list of rules, a cluster of facts and an interpreter.

Rules

The term rules represent what to do or not to do while certain conditions are satisfied. Similarly, domain knowledge is represented by a set of rules. IF First premise, and Second premise, and ... THEN Conclusion The IF side of the rule is referred to as the left hand side (LHS) and the THEN side of the rules referred to as the right hand side (RHS). This is semantically the same as a Prolog rule: Conclusion:- first premise, second premise .

Note that this is confusing since the syntax of Prolog rely on THEN IF, and the normal RHS and LHS appear on opposite sides. Antecedents are evaluated based on what is currently known about the problem being solved. This mean if all antecedents (premise) of the rule evaluated is true, the actions in the consequents part is executed. Each antecedent of a rule typically checks if the particular problem instance satisfies certain conditions. When the consequents of the rules are executed, the rule is said to be fired.

Rule based reasoning techniques

As described in (Sonmenz, 1988). Inference with rules is divided into two, forward chaining and backward chaining.

Backward Chaining: Backward chaining is a goal-driven approach in which you start form an expectation of what is to happen (hypothesis), then seek evidence that supports (or contradicts) your expectation. This entails formulating and testing intermediate hypothesis (or sub hypothesis).

Example: IF Father Diabetes prone THEN Patient diabetes prone.

This would be formulated in Prolog like this**: Diabetes prone (Patient):- Diabetes prone (father (Patient)).**

Forward Chaining: Forward-chaining is a data-driven approach. In this approach we start from available information as it comes in, or from a basic idea, then try to draw conclusions.

Forward chaining, however, would be implemented such that a known fact. Fact = diabetes prone (father (john)).

Advantage of rule based reasoning

Rule based reasoning approach have a numbers of good features. According to (Alechina, 2012) the major advantages of rule based reasoning in the development of knowledge based system are:

- Compact representation of general knowledge. Rules can easily represent general knowledge about a problem domain.
- Homogeneity. Rule based representation has uniform syntax. Hence, the meaning and interpretation of each rule can be easily analyzed.
- Independent. In rule based knowledge representation a new rule can be added without affecting the existing rules. Each rule is an independent piece of knowledge about the problem domain.
- Naturalness of representation. Rules are a very natural knowledge representation method with a high level of comprehensibility. Rules can emulate the expert's way of thinking in natural expression.
- Modularity. Each rule is a discrete knowledge unit that can be inserted into or removed from the knowledge base without taking care of any other technical detail. This characteristic grants flexibility of rule-based reasoning. Because it enables incremental development of the knowledge base.

- Provision of explanations. The ability to provide explanations for the derived conclusions is a straightforward manner. This feature of symbolic rules is a direct consequence of their naturalness and modularity.

Disadvantage of rule based reasoning

As rule based reasoning of prototype knowledge based system has many advantages. But, it has the following limitations as described by (atzilygeroudis, 2007).

- Knowledge acquisition bottleneck- The standard way of acquiring knowledge through interviews with domain experts is bulky and time-consuming.

- Brittleness/fragility of rules- It is not possible to draw conclusions from rules when there are missing values in the input data.
- Inference efficiency problems- In certain cases the performance of the inference engine is not the desired one especially when the rules are too large.
- Difficulty in maintenance of large rules- The maintenance of rule bases is getting a difficult process as the size of the rules increases.
- Interpretation problems- The general nature of rules may create problems in the interpretation of their scope during reasoning process.

2.2.7. Approaches to evaluate the performance of RBR system

Evaluation can be defined as an iterative process of systematic assessment of knowledge based system. The evaluation process carried out at different stage of system development life cycle. The performance of the system was assessed or measured through quantitative and qualitative techniques to achieve the expected objective.

Evaluation must follow an order, it has to be planned and it must be controlled to reduce the cost of the final system (Gomez, 1994).Basically, system performance evaluation procedure tends to assess not only the technical aspects of knowledge based system but also user's satisfaction.

The evaluation criteria depend up on the purposes of the designed knowledge based system in the domain area. Therefore, system's evaluation process tries to evaluate whether a set of rule

achieved the expected goal or not (Lech, 2000).Knowledge based system evaluation process involves to determine the suitability and desirability of the prototype (Mak, 2010). Effective knowledge based system evaluation process incorporates both technical and non-technical aspects. The technical aspects include exploring of the code, examining the correctness of reasoning techniques, checking the efficiency and performance of the system and debugging errors in the early age of a system development. The non-technical aspect includes system compatible with users' satisfaction, the easiness of the system, the quality of the user interface and the acceptability of the system in the real world environments (Seblewongel, 2011).

2.2.8. Methods of evaluation

Knowledge based systems evaluation method can be split into Verification, validation, assessment of human factors and assessment of clinical effect (Thomas, 2001).These evaluation methods are discussed as follows:

Verification: is an evaluation process that should be implemented during system design and development to answer the question 'Did we build the system correctly'. Verification can be defined as the process that involves checking for compliance with the system specifications, checking for syntactic and semantic errors in the knowledge based system. Specification assessment includes user interface, explanation facility, real time performance and security provisions specified in the system design. To verify a knowledge based system, it is possible to use either a program proof or a test strategy. The program proof confirms total correctness of the program logic with mathematical methods and the test proof strategy confirms partial correctness of the program with given test cases (s).

Validation: The concept of validation refers to determining the correctness of the system with respect to users' needs. Validation criteria include comparisons with known results (e.g. past cases or solved problem), comparison against expert performance, and comparison against theoretical possibilities. Empirical validation checks whether the results of content remain stable when the system is under full workload. The system test examines the complete system performance in its working environment. Validation tests include user acceptance surveys, direct comparison on random test cases between human expert and that of the system.

Evaluation of human factors: is the process of determining the acceptability and usability of the knowledge based system. Usefulness of a system is often measured by examining user

satisfaction. User satisfaction measured from different point of views such as content satisfaction, interface satisfaction and institutional objective. Personal aspect such as individuals' dislike of computer takes into consideration.

Evaluation of explanations: Is used to evaluate the explanation ability of knowledge based system in (Agrawal R. &., 1994) explanation facility was measured based on the following criteria which used to judge the user interaction with the system:

- Naturalness- Explanations should be natural to the end user. Explanations that are not structured according to standard patterns of human dialogue create ambiguous information.
- Responsiveness- An explanation facility must have the ability to accept feedback from the user and provide response for the given feedback.
- Flexibility- An explanation facility must be able to offer brief description in more than one way.
- Sensitivity-An explication module should take into account the user's goals, the problem domain and the previous explanatory dialogue.
- Fidelity/accuracy- An explication system must accurately reflect the system's knowledge and reasoning.
- Sufficiency- An explanation module should be able to answer a range of questions that a users' wishes to ask and not limited to those questions predicted by developers.

2.2.9. Knowledge based system development tools

A KBS tool is a collection of software instructions and utilities taken to be a software package designed to support the development of knowledge-based systems. KBS can be built using programming languages such as LISP and Prolog. (McCarthy, 1960) Published a paper showing a handful of simple operators and a notation for functions, one can develop a full programming language. He named this language LISP (List Processing) because one of his main personal views was to use a simple data structure called a list for both code and data. There are several versions of LISP, such as KLISP and C Language Integrated Production System (CLIPS). Prolog is a logic programming general purpose fifth generation (AI) language (U. Nilsson and J. Maluszynsk.i, 2000).It has a purely logical subset, called "pure Prolog", in addition to a number of extra logical features.

Prolog has its roots in formal logic, and in contrast to numerous other programming languages, Prolog is declarative. The program logic is expressed in terms of relations, and execution is activated by running queries over these relations. The language was first believed by a group around Alain Colmerauer in Marseille in the early 1970s. According to (Kowalski, 1988). the first Prolog system was developed in 1972 by Alain Colmerauer and Phillipe Roussel.

As described by (S. Robertson and J. Kingston, 2012), there are around 200 KBS tools and the products are grouped into three main categories based mainly on functionality which also occur to vary markedly in the platforms on which they are available. These groups are: Shells, Languages, and Toolkits. Inference ART and KEE are among the first commercially effective toolkits to build KBS (Akerkar P. S., 2010.). In addition to support towards knowledge acquisition and representational features, there are extra features like price, flexibility, ease of use, user friendliness and vendor availability and support, and documentation support from the tool need to be weighed before final selection.

2.3. Review of Related Works

Previously research attempts in the area of diabetes diagnosis and Advisement have proved applicability and contribution on knowledge based system by different researchers are as follows:

As (K. Rajesh, V. Sangeetha, 2012) suggested that, aimed to apply an Application of Data Mining Methods and Techniques for Diabetes Diagnosis by using a classification C4.5 algorithm, they tested data mining algorithms to predict their accuracy in predicting diabetic status from 8 attributes and 768 record samples and with class variable tested positive or tested negative. The classification rate of 91% was obtained for C4.5 algorithm after they compared different classification algorithms. They recommended that as the future work motivation of the C4.5 algorithms to improve the classification rate to achieve greater accuracy in classification.

As (Thirumal P. C. and Nagarajan N, 2015) suggested that, developed utilization of data mining techniques for diagnosis of diabetes mellitus to check the prevalence of diabetes among young adults and old age people by classification algorithms, Naïve Bayes, Decision tree (c4.5), k-Means, SVM and k Nearest Neighbor are compared and they choose C4.5 algorithm outperforms the other algorithms with the accuracy of 78.25%. Naïve Bayes ranks second with accuracy of 77.8 % and k-NN scores 77.73%, finally SVM acquired 77.4% as accuracy.

As (Govindan, 2012) Suggest that using Rule-Based Reasoning and Object-Oriented Methodologies to Diagnose Diabetes. The study was aimed to diagnosis Type1 and Type2 diabetes depending upon the sign and symptoms using object oriented methodology. The author intends to implement RBR (Rule Based Reasoning) and OO methodologies.

As (Baran Hashemi, 2012) Suggested that to develop An Approach for Recommendations in Self-Management of Diabetes based on Expert System On Hyperglycemia and hypoglycemia cause of the diabetes Making appropriate to give Advisement knowledge about normal blood glucose levels and related signs and symptoms Using methodology multi agent system and Visual C# 2008, aiming at nutrition recommendations applicable in blood glucose self-management. The author intended to raise the complication of diabetes Hyperglycemia and hypoglycemia for managing blood glucose level.

As (Ibrahim M. Ahmed, 2014) Suggested that the Development of an Expert System for Diabetic Type-2 Diet using visual prolog programming language, by using English and Arabic language .the methodologies they used are interviews , document analysis and rule based knowledge representation was used . To provide self-monitor for patient of type 2 diabetes, to get proper n amount of daily calories with list of proper diet satisfies the amount of the calories.

As (Solomon Gebremariam, 2013) Suggested that Design and develop a prototype self-learning knowledge- based system that can provide Advisement for physicians and patients in order to facilitate the diagnosis and Advisement of diabetic patients. For typ1 and type 2 diabetes diseases used rule based knowledge representation techniques and swi-prolog language for implementation of the system, his work was updating the facts. A prototype registered 84.2% performance after extensively tested and evaluated. In this study he recommended for further research to make the integration of data mining techniques with knowledge based system in order to get useful knowledge using data mining techniques.

Therefore the study is going to develop a prototype that integrating predicted data mining result with knowledge based system to advise diabetes patients or experts.

CHAPTER THREE

3. KNOWLEDGE DISCOVERY USING DATA MINING TECHNIQUES

As discussed in chapter two the need of data mining for the knowledge based system is to discover unseen or to extract hidden knowledge to reveal useful knowledge by using data mining techniques like classification, clustering and associations. So using data mining techniques solve the problem of tacit-knowledge acquisition from domain engineers are using traditional knowledge acquisition like interview.

For the purpose of this study classification technique is selected because the Classification divides data samples into target classes. The classification technique predicts the target class for each data points. It is supervised learning approach having known class categories. From the classification models five algorithms are selected for this study these are Jrip, PART rules and J48, LMT and REPTree tree classifiers are selected for experimental.

3.1. Data Understanding and selection

In the activity of data mining process data gathering and understanding of the data is the first step in order to get useful information for the knowledge discovery, The Dataset is collected from Addis Ababa Black Lion Hospital of Diabetes Association Clinic Center. The dataset contains 12,139 instances with 9 attributes (see APPEDIX II.) and two classes Type1 and Type2 diabetic.

Description of the Initial Data

The data collected from this hospital is using the same standard paper format and they are not directly collected for the purpose of this study. Initially the data were in record file format. The data in record file format are then exported to Microsoft Excel for further preparation and ready for as input to WEKA 3.6.9 data mining tool. This initial dataset is described and visualized using Microsoft excel. Then data preprocessing is performed to verify the quality of the dataset such as missing values, error values and to obtain high level information regarding the data mining questions. Table 3.1 shows the data set information which is collected from the hospital.

Table 3-1 Description of initial dataset

Categories	Total
No of Instances	12,139
Number of Attributes	9
Data Size	496KB (507,904 bytes)

As shown in table 3.1 the initial data collected from Black Lion Hospital has 9 attributes. Berry & Linoff (2004) have pointed out that data mining expects data to be in a particular format. First, all the data should be in a single table. Second, each row should correspond to an entity that is relevant to the business. Third, columns with a single value should be ignored. Fourth, columns with a different value for every column should be ignored. Last, for predictive modeling, the target column should be identified and all synonymous columns removed.

One important feature of WEKA, which may be crucial for some learning schemes, is the opportunity of choosing a smaller subset from all attributes. One reason could be that some algorithms work slower when the instances have lots of attributes. Another could be that some attributes might not be relevant. Both reasons lead to better classifiers. As a result of the dataset from hospital as business process analysis relevant attributes are selected in the subsequent sections for instance the patient id, card number, date of the patient diagnosed, the name of the patients are not used for this research after discussing with domain experts and this eight independent variables are the most important to distinguish the types of diabetes, Moreover, data mining techniques such as discretization or categorization is used to summarize the low level too many distinct values into high level values. The following Figure 3.1 shows that the selected attributes for this study after discussed with domain experts.

	Year	Sex	age	FBS	FamilyHistory	Obesity	blurred-vision	ketone	weight-loss	Type
2	2012	M	52	121	yes	no	no	no	no	2
3	2012	M	66	93	yes	yes	no	no	no	2
4	2012	M	62	114	yes	no	yes	no	no	2
5	2012	F	56	181	yes	no	no	no	no	2
6	2012	F	29	210	no	no	no	yes	no	1
7	2012	M	33	244	yes	no	no	no	yes	1
8	2012	M	68	168	yes	yes	yes	no	no	2
9	2012	M	50	223	yes	yes	no	no	no	2
10	2012	M	49	240	yes	yes	yes	no	no	2
11	2012	F	21	111	no	no	no	yes		1
12	2012	M	49	234	yes	yes	yes	no	no	2
13	2012	M	33	313	yes	yes	no	no	no	2
14	2012	M	31	89	no	no	no	yes	yes	1
15	2012	M	26	234	yes	no	no	no	no	2
16	2012	F	17	171	no	no	no	yes	yes	1
17	2012	M	33	211	no	no	no	yes	no	1
18	2012	M	78	212	yes	yes	no	no	no	2
19	2012	F	65	234	no	yes	no	no	no	2
20	2012	M	23	111	no	no	no	yes	yes	1
21	2012	M	67	213	yes	yes	yes	no	no	2
22	2012	F	16	156	no	no	no	yes	yes	1
23	2012	M	45	234	yes	no	no	no	no	2
24	2012	F	23	123	no	no	no	yes	no	1
25	2012	M	59	341	yes	yes	no	no	no	2
26	2012	M	41	125	yes	yes	no	no	no	2

Figure 3-1 Data set before categorization of attributes

3.2. Data Preprocessing

This step is removing noise and inconsistent data after the data is loaded into weka tool; combining multiple data sources, filling the missing values, and pruning is the most important task during data preprocessing. The data set collected is first in the form of CSV (Common Separated Value), and then the data was converted into the ARFF format (Attribute Relation File Format) for experimentation by using Weka 3.6.9 tool. The following figure shows that the preprocessing panels applied on the datasets.

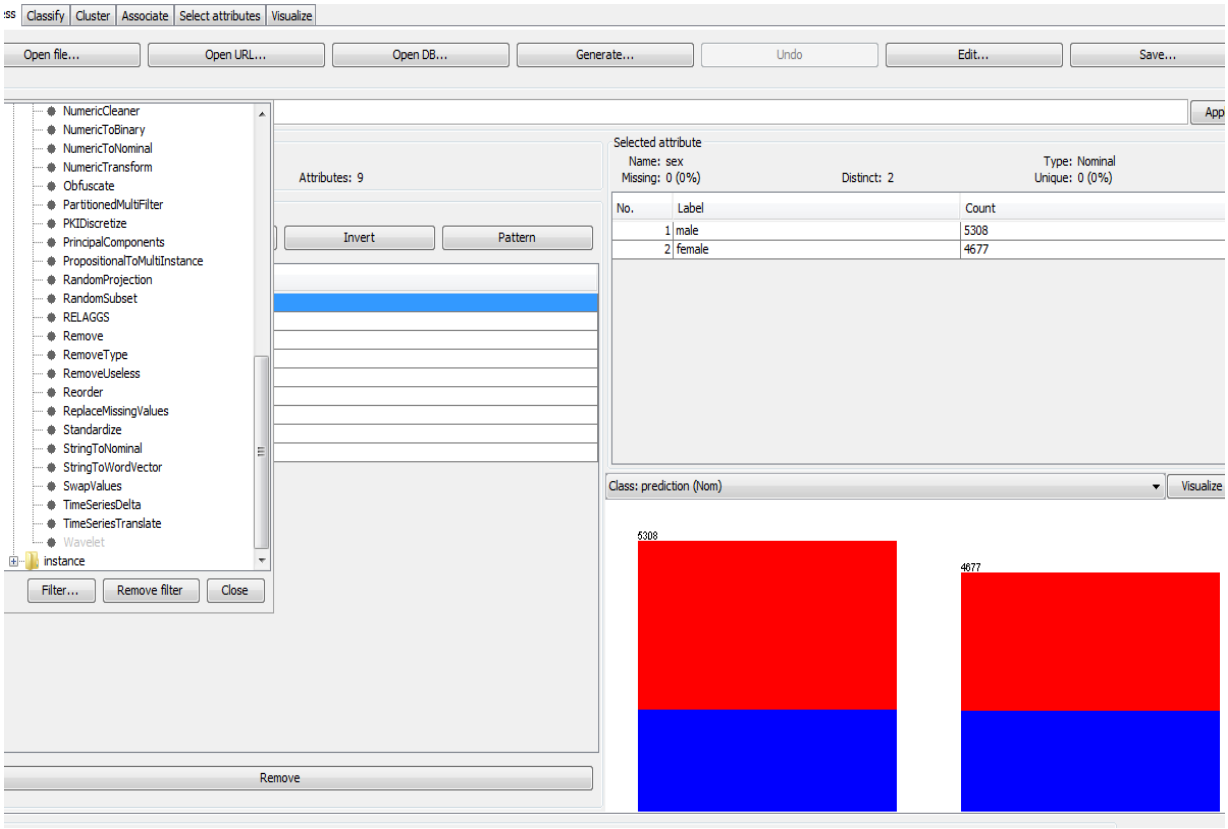

Figure 3-2 Data Preprocessing Panel

From figure 3.2 we observe that it is possible to apply different data preprocessing methods under filters unsupervised or supervised techniques by choosing attribute or instances property of the data set in order to fill missing values and discretize the data , for this study ReplaceMissingValues filter was used to preprocess the dataset see Table 3.2 below.

Table 3-2 Lists of some filters

Discretize	An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes
NominalToBinary	Converts all nominal attributes into binary numeric attributes
Normalize	Scales all numeric values in the dataset to lie within the interval [0, 1]
ReplaceMissingValues	Replaces all missing values for nominal and numeric attributes with the modes and means of the training data
StringToWordVector	Converts a string attribute to a vector that represents word occurrence frequencies

3.2.1. Data transformation

In data transformation, the data are transformed or consolidated into forms appropriate for mining. For example, smoothing techniques including binning, regression and attribute construction are the most used ones. From the dataset the "AGE" and "FASTING BLOOD SUGER" attributes were discretized (binned) to reduce the distinct values of the attributes so that it will suit the mining tool and to obtain meaningful patterns. Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels there by reduces and simplifies the original data. This leads to a concise, easy to use, knowledge-level representation of mining results. A concept hierarchy for a given numerical attribute defines a discretization of the attribute (Han & Kamber, 2006). Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as the numerical

values for the attribute "AGE") and high-level concepts (child, young, and old). Therefore, the researcher performed discretization on the attributes "AGE" and "FASTING BLOOD SUGER" using a binning method. The attribute "AGE" is binned into three levels (**child, young**, **adult** and **old**) whereas the attribute "FASTING BLOOD SUGER" is binned to **low**, **medium** and **high.**

Categorizing value of Age Attribute

For easy interpretation or to prepare the data for mining purpose it is important to categorize the data into the same data types. The attributes of the datasets which are collected from the sources are in different data types. Due to this, it is important to convert or transform the data into understandable way for data mining purpose. For instance the age attributes in the data set has the range values of below 15, 16-35, 36-59 and greater than or equal to 60. Since the researcher decided to categorize the age as described by (Petry, 2002) " Child", "Young" "Adult" and "old". Table 3.3 shows age categorization.

Table 3-3 Age Attribute Categorization adopted from (Petry, 2002)

Age	Category
Below 15	Child
16-35	Young
36-54	Adult
Greater than or equal to 55	Old

Categorizing value of Fasting Blood Sugar Attribute

This attributes the measure of the diabetes blood sugar in order to identify the patients' Blood sugar level, this attributes also recorded in the dataset as integer data types with the range of below 100 mg/dl 100-300mg/dl and greater than 300 mg/dl. Since the researcher try to categorize this attribute ranges after asking the experts how they give the name for the ranges of this attribute in order to identify diabetes types with the combination of other attributes. Therefore it is categorized as: "low", "Medium", and "High" as shown in the following table 3.4

Table 3-4 Categorizing value of Fasting Blood Sugar Attribute

Fasting Blood Sugar	Meaning
below 100mg/dl	Low
100-300mg/dl	Medium
>=300 mg/l	High

So far , different attributes are transformed into their appropriate values for data mining purpose in order to predict the correct class of diabetes in addition to this, the researcher transforms the attributes like blurred-vision, family history (genetic existence), Ketone, Obesity and Weight-loss attributes into nominal values (yes, no) in order to simplify the complexity of different data types.

3.3.WEKA Format Dataset

After all, the next important issue was preparing the selected dataset in WEKA understandable format (ARFF and CSV) for experimentation. The data were initially in a spreadsheet format (xlsx extension), the researcher convert from a spreadsheet to ARFF. And this is done by saving spreadsheet format as comma separated value (CSV) format then opened the CSV file format in WEKA explorer and save it as an ARFF file as shown in listing 3.2.1. The Header of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types. Lines that begin with a % are comments. The @RELATION, @ATTRIBUTE and @DATA declarations are case insensitive. The ARFF format merely gives a dataset; it does not specify which of the attributes is supposed to be predicted. This means that the same file can be used for investigating how well each attribute can be predicted from the others or it can be used to find association rules or for clustering.

```
@relation Diabetes
@attribute sex{male,female}
@attribute age {child,young,adult,old}
@attribute fbs{low,medium,high}
@attribute blurred_vision{yes,no}
@attribute family_history {yes,no}
@attribute obesity {yes,no}
@attribute ketone {yes,no}
@attribute weight_loss {yes,no}
@attribute prediction {type1,type2}
@data
male,adult,high,yes,yes,yes,no,no,type2
female,young,medium,no,no,no,yes,yes,type1
male,adult,high,yes,yes,yes,no,no,type2
male,young,high,no,yes,yes,no,no,type2
male,adult,high,yes,yes,yes,no,no,type2
female,young,medium,?,no,no,yes,yes,type1
```

Listing 3-1 Sample ARFF Files of Diabetes Data

3.4. Building classifiers

After the input dataset is transformed in the format that is suitable for the machine learning scheme, it can be fed to it. Building or training a classifier is the process of creating a function or data structure that will be used for determining the missing value of the class attribute of the new unclassified instances. The concrete classifier can be chosen from the 'Classify' panel of the Explorer. Data mining techniques are used to build a model according to which the unknown data tries to identify the new information. Commonly used techniques for data classification and prediction in data mining are decision tree induction and rule-based induction. The models were built with five different supervised machine learning algorithms i.e. three Decision Tree Classification Algorithms and two rule based classifiers. These algorithms have the ability to produce rules which is crucial for understanding the hidden knowledge.

Table 3-5 A list of classifiers experimented

Classifiers	Descriptions
J48tree	Class for generating a pruned or unpruned C4.5 decision tree
Jrip	This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER.
PART	Class for generating a PART decision list. Uses separate-and-conquer
LMT	Classifier for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves.
REPTree	Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). Only sorts values for numeric attributes once.

3.5. Data Mining Model

In this step five appropriate data mining classification model or algorithms are ready to apply on the dataset. The classifier techniques which are selected for this study are, JRip, PART, J48, LMT and REPTree based on their accuracy the best algorithms was selected for discovering the knowledge.

Experiment on JRip Classifier

JRip (RIPPER) is one of the basic and most popular algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error, JRip (RIPPER) proceeds by Advisementing all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the members of that class. It classified instances correctly 92.6216% and incorrectly classified instances **7.4 %.** It generates **52 rules** and takes **5.62** seconds time to build model.

Experiment by PART Classifier

Class for generating a PART decision list uses separate-and-conquer. It builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule. This algorithm is also classifier rule **induction** which generates the rules. For this experiment also 10-fold-cross-validation is applied and the accuracy of this experiment is **93%** correctly classified and **7.2 %** instances are incorrectly classified. By this experiment **65** rules are generated in **1.37**seconds.

Experiment by J48tree Classifier

It is an algorithm used to generate a decision tree and is an extension of Quinlan's earlier ID3 Algorithm. The decision trees generated by this can be used for classification and so referred to as statistical classifier. In this experiment 10-fold-cross-valiadtion is used .It correctly classified instances with **93.84 %** and incorrectly classified instances **6.16 %** of accuracy respectively. From the experiment **55 numbers** of leaves and **110** Size of the tree are generated in **1.31** seconds**.**

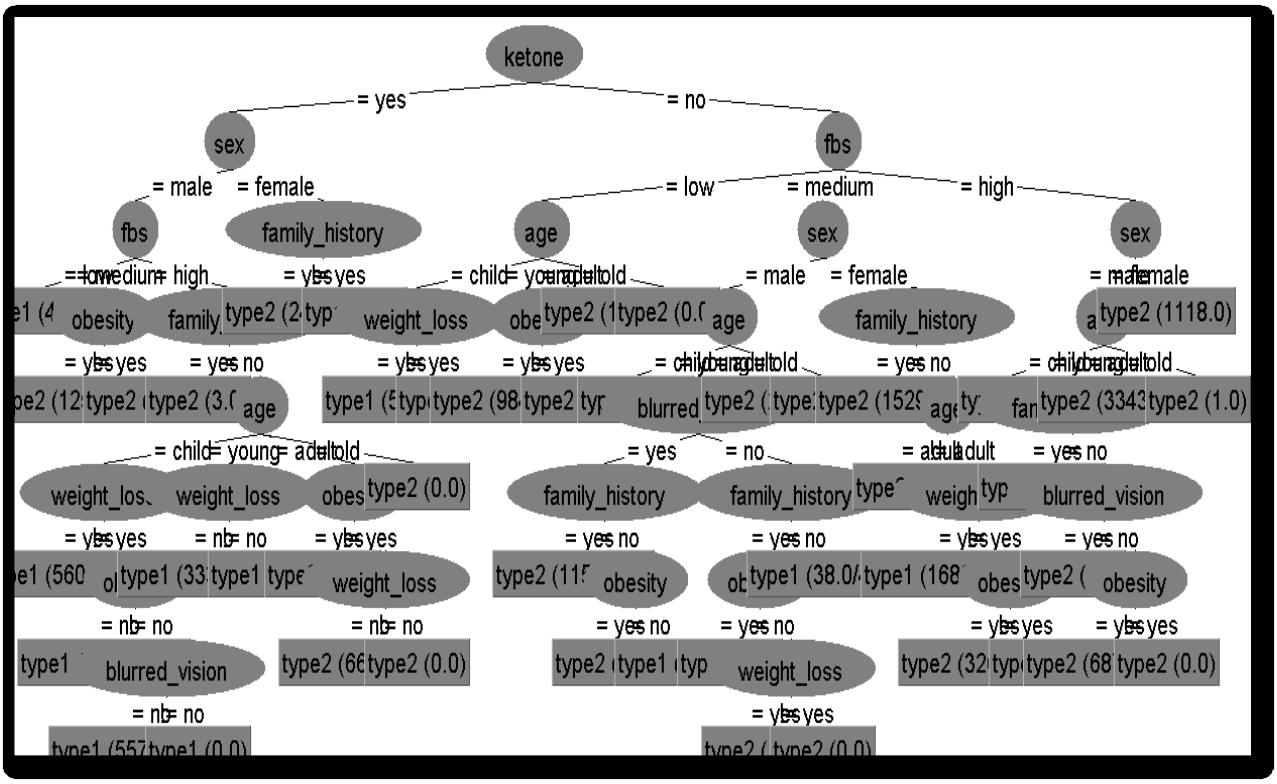

Figure 3-3 Sample Decision tree of j48

Experiment by LMT Classifier

Classifier for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values. 10-fold-cross-validation is also used , The experiment result of LMT algorithm gets 8897 and 685 correctly classified and incorrectly classified instances with the accuracy of 92.8 %and 7.2 % respectively, From The experiment **43** Number of Leaves and **85** Size of the tree are generated in **46.85 seconds**

Experiment by REPTree Classifier

REPTree is a regression tree representative algorithm, the result of this experiment gives 8892 correctly classified and incorrectly classified instances 690, with **93%,** 7.2 % accuracy respectively, and the tree gives **93** size of tree in **0.42** seconds

3.6. Evaluation methods of models

After building a model, the researcher must evaluate its results and interpret their significance. Remember that the accuracy rate found during testing applies only to the data on which the model was built. In practical, the accuracy may vary if the data to which the model is applied differs in important and unknowable ways from the original data. More importantly, accuracy by itself is not necessarily the right metric for selecting the best model. We need to know more about the type of errors and the costs associated with them. To select the best classifier model from all the experimental there are a lot of mechanisms like confusion matrix, performance accuracy based on correctly classified and incorrectly classified, F-Measure, Recall, Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error etc.

From the following table, 3.6 we can see that, each model registered different performance related to different parameters, like correctly, incorrectly classified, absolute mean error, root mean square error, relative absolute error and time taken to build models. From all algorithms J48 tree classifier registered high performance **(93.84%)** than others. Therefore due to performance score, the rules generated by j48 classifier shall be used for the knowledge base system development.

Table 3-6 Stratified cross-validations Summary of each classifier

Classifiers	Correctly classified	Incorrectly classified	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Time taken to build model
Jrip	92.6216 %	7.3784 %	0.8467	0.1178	0.2429	24.5251 %	49.5672 %	5.62 seconds
PART	92.8303 %	7.1697%	0.8513	0.1034	0.2278	21.5293 %	46.4845 %	1.37seconds
J48	**93.84%**	**6.16 %**	**0.1032**	**0.1028**	**0.1005**	**20.2345 %**	**45.2956 %**	**0.65 seconds**
LMT	92.8512	7.1488 %	0.8517	0.2285	0.1006	23.7823 %	47.6244 %%	86.85 seconds
REPTree	92.8199 %	7.1801 %	0.1035	0.2279	0.2279	21.5536 %	46.5024 %	0.08 seconds

Confusion Matrices

For classification problems, a confusion matrix is a very useful tool for understanding results. A confusion matrix (Table 3.7) shows the counts of the actual versus predicted class values. It shows not only how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong. The rows of the confusion matrix represent the actual class of the test cases; the columns represent what our classifier predicted. The following tables are a confusion matrix of J48 tree, JRIP, PART, LMT and REPTree results. The columns show the actual classes, and the rows show the predicted classes. Therefore the diagonal shows all the correct predictions.

In the following confusion matrix of five models we can see that the J48 model predicted 3543 of the Type1 correctly, but misclassified 383 of them: as Type2. This is much more informative than simply telling us an overall accuracy rate of **92.3%** (3543 correct classifications out of 383 cases) the total accuracy of each class. So far, we have been evaluating our classifier by computing the percent accuracy.

Table 3-7 Confusion Matrices of each classifier

		ACTUAL	
CLASSIFIERS	**PREDICTION**	Type1	Type2
J48tree	Type1	3543	301
	Type2	383	5355
Jrip	Typ1	3510	330
	Type2	377	5365
PART	Type1	3540	300
	Type2	387	5355
LMT	Type1	3541	299
	Type2	386	5356
REPTree	Type1	3540	300
	Type2	388	5354

Accuracy by Class

This comparison method shows accuracy of each classes depending up on the parameters like, True Positive Rate, False positive Rate, Precision, Recall, F-measure , ROC and Weighted Avg of each classes, as shown in the Table 3.7. The values of each class are almost similar in numbers for each Algorithm. But FP rate of the algorithms for the classes shows some differences; due to these J48tree classifier algorithms shows less number of FP rate .therefore the researcher select this classifier algorithm.

Table 3-8 Accuracy by TP, FP Rate, Recall and F-Measure of each Class

Classifiers	True positive	False positive	Recall	F-Measure	Type
Jrip	0.914	0.066	0.914	0.909	Type1
	0.934	0.086	0.934	0.938	Type2
PART	0.922	0.067	0.922	0.912	Type1
	0.933	0.078	0.933	0.94	Type2
J48	0.922	0.067	0.922	0.911	Type1
	0.933	0.078	0.933	0.94	Type2
LMT	0.922	0.067	0.922	0.912	Type1
	0.933	0.078	0.933	0.94	Type2
REPTree	0.922	0.068	0.922	0.911	Type1
	0.932	0.078	0.932	0.94	Type2

3.7. Selecting Best Model

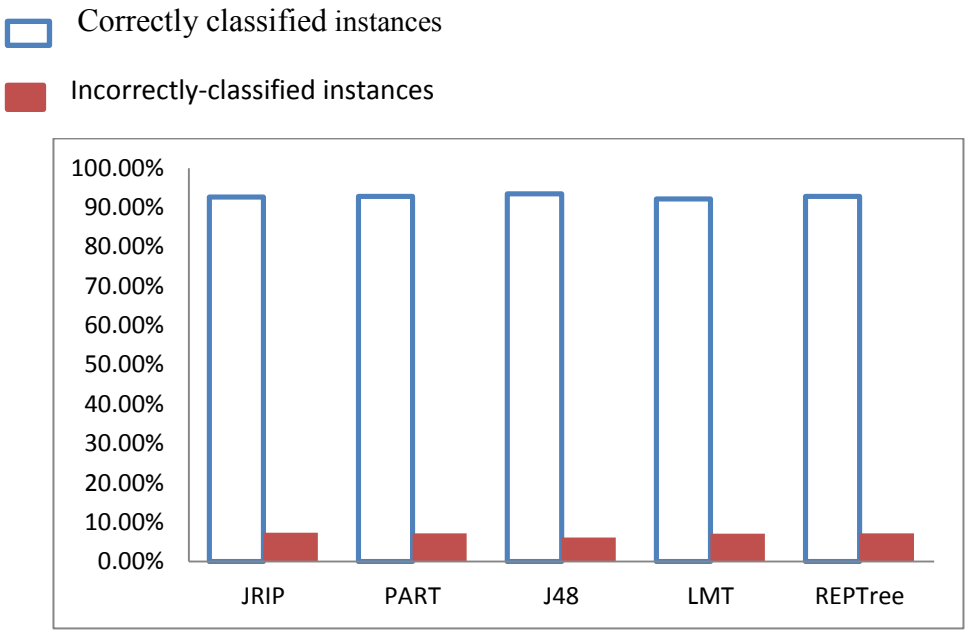

Figure 3-4 Predicted Accuracy of each model by cross-validation

From figure 3.3 the researchers observe that j48 classifier tree model has correctly classified the data than others with **93.84%** accuracy and **6.16%** incorrectly classified, the other classifiers are almost registered similar accuracy**.** Therefore for this study the J48 classifier algorithm is selected. It registers good performance values or accuracy compared with other algorithms. Table 3.11 shows partial sample rules generated by J48tree.

Table 3-9 Partial list of Rules generated by J48

Rule1: IF ketone = no AND FBS = high AND age= old THEN Type2 (420.0)
Rule 2: IF Ketone = no AND FBS = medium AND Famliyhistory = yes THEN Type2 (3.0)
Rule3: IF ketone = no AND FBS = High NAD Famliyhistory = yes THEN Type2 (165.0)
Rule4: IF Famliyhistory = yes AND Obesity = yes THEN Type2 (2438.0)
Rule 5: IF weight loss = yes AND ketone = yes NAD obesity = no THEN Type1 (234.0)
Rule 6: IF Age = Child AND ketone = yes AND Obesity = no THEN Type1 (330.0)
Rule 7: IF Family history = yes AND obesity = no AND age = Child THEN Type2 (482.0)
Rule 8: IF Age = Adult AND Obesity = yes AND Weight loss = no THEN Type2 (34.0/2.0)
Rule 9: IF FBS= High AND Age = Adult AND Obesity = yes THEN Type2(23.0/3.0)
Rule 10: IF FBS= High AND Obesity = yes: type2 (1928.0/229.0)
Rule 11: IF FBS = Medium Obesity = no AND family history = yes: type2 (482.0)
Rule 12: IF Age = Adult THEN Type2 (253.0)
Rule 13: IF famliyhistory = yes AND Obesity = yes THEN type2 (3.0)
Rule 14: IF famliyhistory = yes AND Obesity = no THEN type2 (67.0)

CHAPTER FOUR

4. KNOWLEDGE ACQUISITION, MODELING AND REPRESENTATION

Knowledge acquisition is the process of extracting, structuring and organizing knowledge from human experts and other sources such as books, databases, the internet, research papers, documents, one's own experience and transferring to the knowledge base (Sandoval, 1995). It is a prerequisite and an important phase of knowledge-based system development that deals with the activities of capturing knowledge from human experts, books, documents or computer files. It is used to develop a computational problem solving model, specifically a program to be used in some consultative or advisory role (Boose, 1988). The knowledge which is captured in the knowledge acquisition process may be specific to the problem domain or to the problem solving procedures according to (Int. J. Man, 1987).

Acquisition of knowledge is the most important and decisive phase in building Knowledge-Based Systems. However, it is an extremely hard task; capable of making an error, that knowledge engineer does while developing a Knowledge-Based System. As a result of the challenges and difficulties confronted in the transfer of expertise knowledge, knowledge acquisition has been depicted as the obstruction of Knowledge-Based Systems development. Accordingly, this study gathers and analyzes the basic domain knowledge regarding diabetes advisement and the essential concepts which are identified about the diabetes managements and that help to develop the proposed prototype knowledge-based system. The knowledge engineer collects diabetes managements/recommendation measure and models it by using decision tree structure. The study explores the applicability of rule-based reasoning. The knowledge for this study is acquired from domain experts by using interviewing and critiquing knowledge elicitation methods and from relevant documents by using documents analysis technique which has been employed to purify the acquired knowledge for the advisement purpose.

4.1. The knowledge Acquisition Process

For the purpose of this research, the process of knowledge acquisition includes some basic activities such as interview of domain expert's and review of relevant sources of information. The objective of knowledge acquisition is to gather the required knowledge, interpreting the acquired knowledge, analyzing and validating the knowledge content. Based on the acquired knowledge, the proposed knowledge based system is designed using decision tree model. Therefore, knowledge acquisition process of this thesis is based on domain expert interviewing of and reviewing of related documents. This section discusses the detail of knowledge acquisition techniques as follows.

Interviewing Domain Experts

Both organized and non-formal interviews were employed to elicit tacit knowledge from domain experts. Since one of the specific objectives of this research is extracting tacit knowledge which is entrenched and personalized in experts mind. For this research, four experts were selected purposively for interview about diabetes. These four experts were selected for extensive discussions using structured and unstructured interview to discover relevant tacit knowledge. Table 4.1 summarizes about experts.

Table 4-1 Domain expert's profiles

No	Educational level	Specialization	Roles
1	MD	Medicine	Medical Director
2	MD	Medicine	General
3	MD	Medicine	General
4	MD	Medicine	General

MD= Medical

Table 4-2 General Comparison of type 1 and type 2 diabetes mellitus, Adapted from (Daca, 2004).

	Type1	Type2
Previous terminology	IDDM	NIDDM
Relative frequency	5-10% of the diabetic population	90-95% of the diabetic population
Age at onset	<35 years	Usually >35years
Precipitating and associated risk factors	Largelyunknown; microbial, chemical, dietary etc	Age, obesity, sedentary lifestyle, previous gestational diabetes
Endogenous insulin reserve	Low or absent	Usually present
Stress, withdrawal of insulin	Ketoacidosis	Nonketotic hyperosmolar stat
Human leukocyte antigen association	Positive	Negative
Family history of Diabetes mellitus	Infrequent	Frequent
Associated auto- immune diseases	Yes	No
Clinical picture	Weight loss, Polyuria, Polydipsia	Obese, Strong family history type 2 diabetes, Ethnicity – high prevalence populations
Advisement	Insulin	Diet, exercise, oral hypoglycemic drugs, may be insulin needed

4.2. Diabetes Management and Advisement

Advisement

To try to normalize insulin activity & blood glucose levels in an attempt to reduce the development of the vascular & Neuropathic complications. There are five components of management for diabetes (Daca, 2004).These are divided into two; **Drug-therapy Advisement** and **non-drug therapy Advisement.**

1.Non-Drug-Therapy (Medication**) Advisement.**

Diet Advisement: Diet is a basic part of management in every case. Advisement cannot be effective unless adequate attention is given to ensuring appropriate nutrition**.**

Food items the diabetic should avoid (rapidly absorbed carbohydrates) Sugar, honey, jams, candy, marmalade Cakes, Sweet Biscuits Soft drinks (Coca Cola, Miranda etc) alcohols

Foods which the diabetic can take with restrictions Food items from grains: enjera, bread, kinche, kita, atmit Foods items prepared from peas, beans, lentils, chick peas Potato, sweet potato, kocho, bulla .

Food items the diabetic can take freely or with minimal restrictions : Lean meat and fish Eggs, milk, cottage cheese Green leafy vegetables (cabbage, tomato, pumpkin, carrots, onion) Tea, coffee and lemon juice without sugar, Ambo water, other mineral waters Spices: pepper,

garlic, 'berbere'.

Education: Education of the person with diabetes is an essential component of management in every case. To ensure appropriate management, the basic knowledge and skills should be acquired by the patient and his family and the health care team should work closely with the patient to achieve this objective and to promote self-care. The person with diabetes should also be involved in setting therapeutic targets for weight, blood pressure and blood sugar control.

Exercise: Physical activity promotes weight reduction and improves insulin sensitivity, thus lowering blood glucose levels. Together with dietary Advisement, a programme of regular physical activity and exercise should be considered for each person. Such a programme must be tailored to the individual's health status and fitness. People should, however, be educated about the potential risk of hypoglycemia and how to avoid it. Individuals with type 1 DM are prone to either hyperglycemia or hypoglycemia during exercise, depending on the pre-exercise plasma glucose, the circulating insulin level, and the level of exercise-induced catechol amines. If the insulin level is too low, the rise in catec holamines may increase the plasma glucose excessively, promote ketone body formation, and possibly lead to ketoacidosis. If the circulating insulin level is excessive, this relative hyperinsulinemia may reduce hepatic glucose production (decreased glycogenolysis, decreased gluconeogenesis) and increase glucose entry into muscle, leading to hypoglycemia.

To avoid exercise-related hyper- or hypoglycemia, individuals with type 1 diabetes should monitor blood glucose before, during, and after exercise, delay exercise if blood glucose is > 250 mg/dL, <100 mg/d), or if ketones are present, eat a meal 1 to 3 hours before exercise and take supplemental carbohydrate feedings at least every 30 min during vigorous or prolonged exercise decrease insulin doses (based on previous experience) before exercise and inject insulin into a non-exercising area (Daca, 2004).

2.Drug-Therapy(Medications) Advisement

Insulin therapy: In type 1 diabetes, the body loses the ability to produce insulin, thus, exogenous insulin must be administered. A standard insulin Advisement consists of one or two injection/day, using intermediate or long acting insulin with or without regular insulin. In type 2 diabetes, insulin may be necessary on a long term basis to control glucose levels if diet and oral agents have failed. In addition, some patients whose type 2 diabetes is usually controlled by diet alone or diet and an oral agent may require insulin temporarily during illness, infection, pregnancy, surgery or some other stressful events (Daca, 2004).

Type 1 Patients with type 1 diabetes have on absolute requirement of insulin for survival. Insulin is also used in type 2 diabetics when a combination of oral agents fails to achieve glucose targets and temporarily in patients with serious infection or surgery. The types of insulin available are rapid acting, short acting, intermediate acting and long acting.

Standard insulin therapy consists of one to two injections per day using intermediate or long acting insulin with or without regular insulin. Starting insulin does vary from 0.15 to 0.50/kg (as high as 1.5 u/kg in cases of severe insulin resistance) depending on patient size and degree of glycaemia. Adults of normal weight may be started with 20-25 u/d of intermediate acting insulin and increased to maintain a blood sugar level of 100-120 mg/dl.

Type 2: Provided pharmacologic therapy is not required immediately all patients should be given at least a one month trial of diet, exercise and weight management. If this regimen does not lead to adequate blood glucose control, oral antihyperglycemic agents with or without insulin are indicated. Insulin may be needed in symptomatic patients who have type 2 with FBG (Fasting Blood Glucose) values are greater than 250 mg/dl.

4.3.Conceptual Modeling

Conceptual modeling helps to clarify the structure a knowledge-intensive business task. The knowledge model of an application provides a specification of the data and knowledge structure for the application. (Schreiber, 2011).

According to (Richard et, 2010) one of the most extensively applied methods of conceptual modeling is called decision tree. Decision tree commonly acts as a key role in a knowledge modeling process. Decision tree is used for the search space of a certain problem and presented by a graph. A node in the tree denotes a decision to be attained when finding a solution of a certain problem, and the branches extended from the node show the potential values of the decision. To find the solution of a certain problem, anyone then traces by way of its tree using data of a certain problem to select a branch at every node. Therefore figure 4.1 shows the concepts of diabetes which are acquired from the domain experts to model knowledge based system for diagnosis and Advisement of diabetes in addition to Data mining knowledge discovery methods through data mining prediction method.

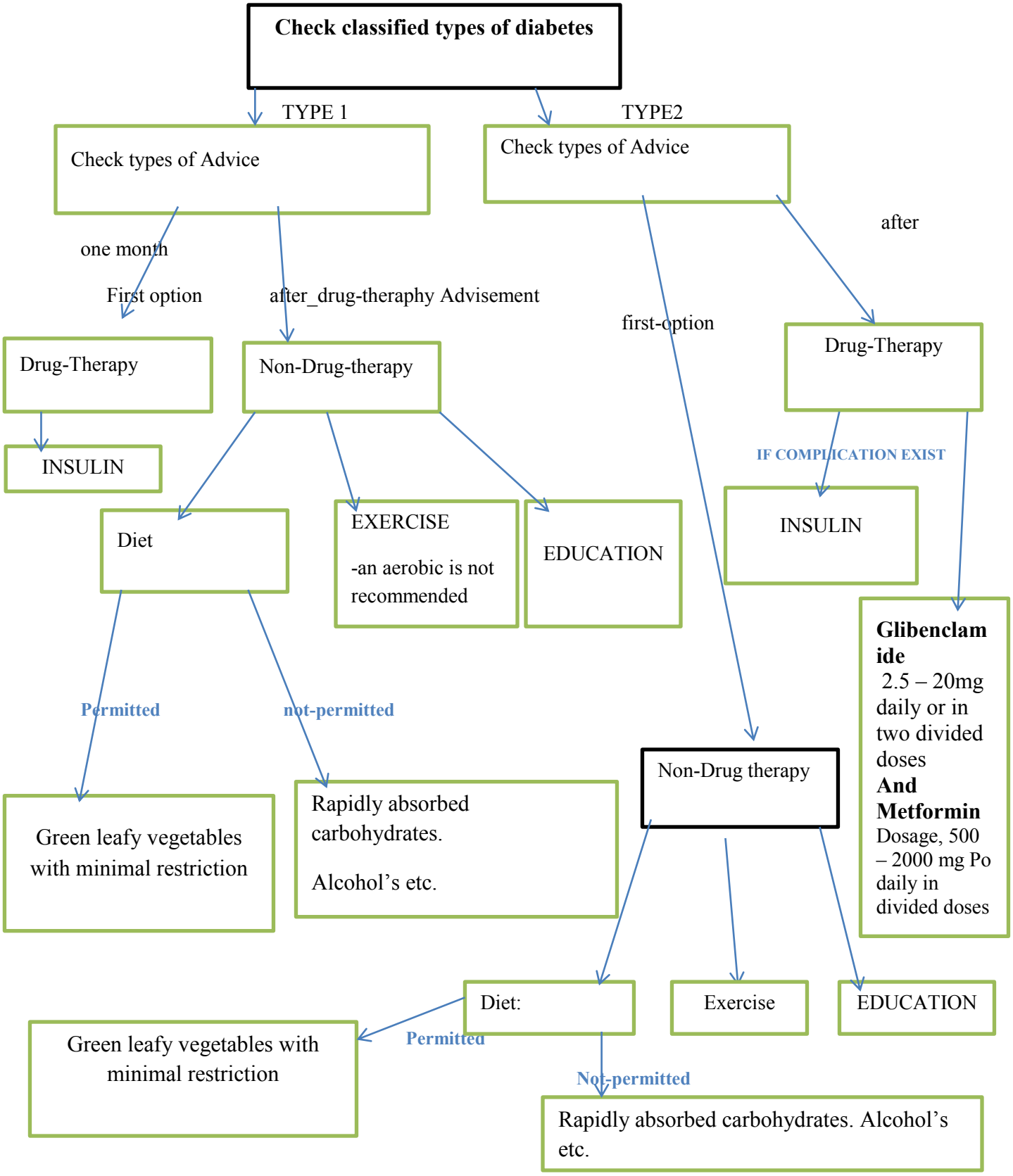

Figure 4-1 Decision tree of diabetes Advisement after prediction rule representation

There are a lot of knowledge representation methods like Frame and Semantic knowledge representation .From those Production, rule systems is the rule representation based on the general underlying idea of condition-action pairs (also called if-then pairs, production rules, or just plain productions).A production rule is written in the form "if this condition holds, then this action is appropriate". They are capable of modeling any computable procedure. The following rules are constructed from the above decision tree after the conceptual model is acquired from the domain experts and data mining results are identified as type1 or type2 diabetes.

Rule1: IF Prediction type ="Type1 diabetes"

Then Advisement =" non-drug therapy is recommended ".

Rule2: IF Prediction type ="Type1 diabetes"

Then Advisement = "Insulin is recommended as the first option".

Rule3: IF Prediction type ="Type2 diabetes"

Then Advisement =" non-drug therapy is recommended ".

Rule4: IF Prediction type = "Type2 Diabetes "AND Complication exist AND diet and exercise is not control blood glucose Then Advisement = "insulin is recommended".

Rule5: IF Prediction type = "Type2 Diabetes "AND no Complication exist AND diet and exercise is not control blood glucose Then Advisement = "**Glibenclamide and Metformin is** recommended".

CHAPTER FIVE

5. Integration of Data Mining Results with Knowledge Based System

The main goal of this study is to integrate data mining result for the development of knowledge based system and to predicting diabetes types based on the selected attributes. The data mining result is the predicted or output which is generated by j48tree as type1 or type2 diabetic based on different attributes used for this study while the classifier algorithms are applied on the training dataset.

The integration system contain two parts: Data mining part which predicts the diabetes types from diabetes dataset based on the selected model and the second one is knowledge based system which acquire the result of data mining by asking the users to answer the attribute values of the diabetes diagnosing method through the user interface .Then jpl_call library loads java calling method to access generated rules from weka data mining tool which embedded in java net bean and giving the Advisement or Advisement based on the diabetes predicted type.

5.1. System Design And Architecture For Integration

Figure 5.1 below shows the overall activities of the integration of hidden knowledge about the diabetes prediction and Advisement with the knowledge based system.

.

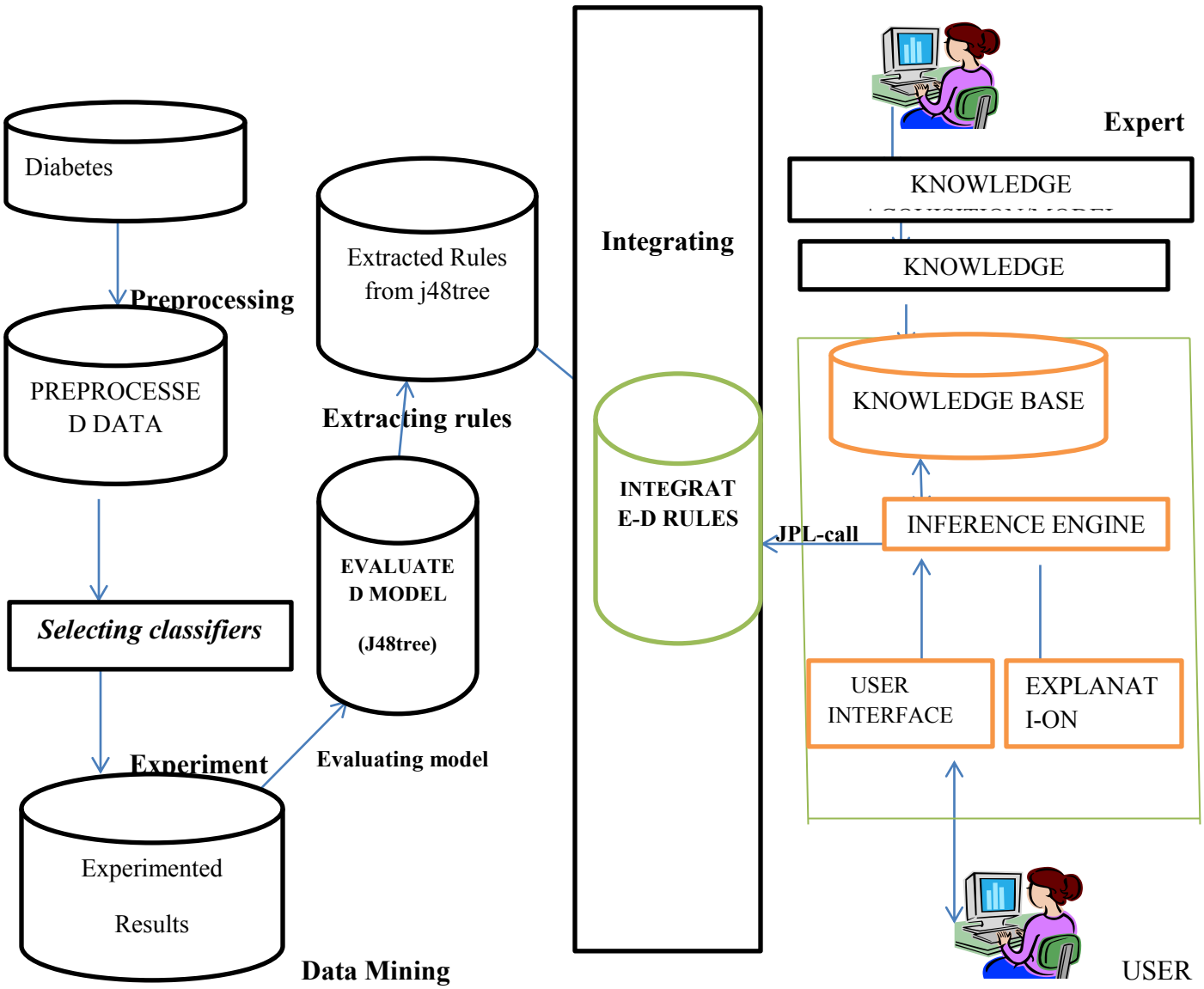

Figure 5-1 Architecture of Integration Data Mining Result with KBS

Extracting the Rules

So, far after the model is selected based on different selection criterion, the selected model is ready to generate the best rules classified. Weka tool is developed by using java program .Therefore j48 tree can generate java source code from the inputted dataset of diabetes; the following step shows how to generate source code of java from j48 classifier:

Start weka tool->import data set->Choose Classifier->select Trees ->select J48->select more Options->classifier evaluation options->check output source code->click Ok->start.

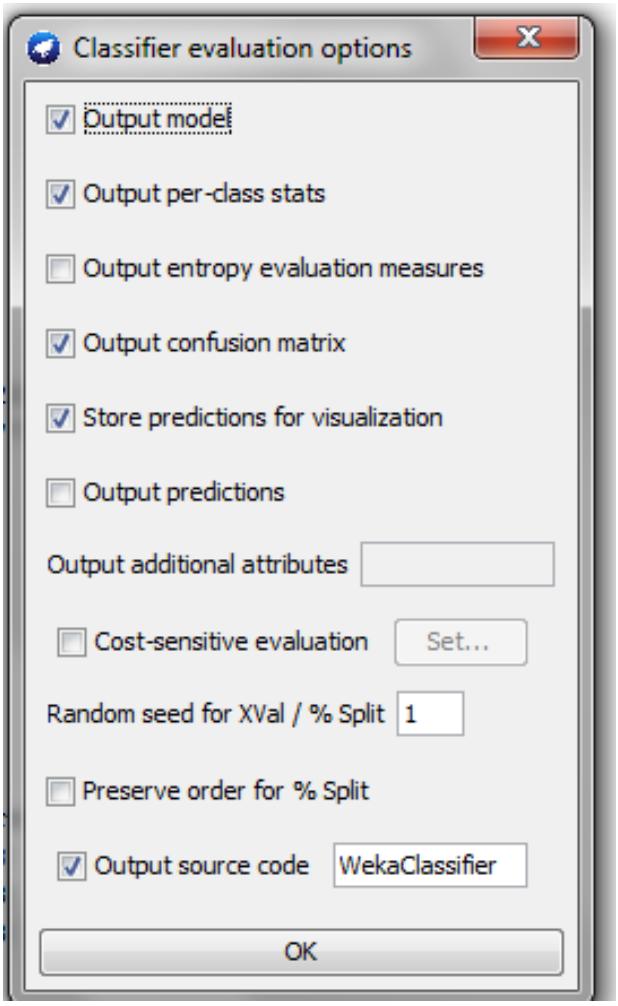

Figure 5-2 Extracting java source code from weka

Here is the partial source code of j48tree generated in weka as java source code.

```
public static double classify(String [] i)
  throws Exception {

  double p = Double.NaN;
  p = WekaClassifier .N8386930(i);
  return p;
 }
 static double N8386930(String []i) {
  double p = Double.NaN;
  if (i[6] == null) {
   p = 0;
  } else if (i[6].equals("yes")) {
  p = WekaClassifier .N1843f4e1(i);
  } else if (i[6].equals("no")) {
  p = WekaClassifier .N36f0c914(i);
  }
  return p;
 }
}
 static double N118a3c216(String []i) {
  double p = Double.NaN;
  if (i[1] == null) {
   p = 1;
  } else if (i[1].equals("child")) {
   p = 1;
  } else if (i[1].equals("young")) {
  p = WekaClassifier .Nccfd6817(i);
  } else if (i[1].equals("adult")) {
   p = 1;
  } else if (i[1].equals("old")) {
   p = 1;
  }
  return p;
 }
 static double Nccfd6817(String []i) {
  double p = Double.NaN;
  if (i[3] == null) {
   p = 1;
  } else if (i[3].equals("yes")) {
  p = WekaClassifier .N1affa8018(i);
  } else if (i[3].equals("no")) {
  p = WekaClassifier .N86942f20(i);
  }
  return p;
```

```
  }
  static double N1affa8018(String []i) {
    double p = Double.NaN;
    if (i[4] == null) {
      p = 1;
    } else if (i[4].equals("yes")) {
      p = 1;
    } else if (i[4].equals("no")) {
    p = WekaClassifier .N3a391919(i);
    }
    return p;
  }
  static double N3a391919(String []i) {
    double p = Double.NaN;
    if (i[5] == null) {
      p = 1;
    } else if (i[5].equals("yes")) {
      p = 1;
    } else if (i[5].equals("no")) {
      p = 0;
    }
    return p;
  }
  static double N86942f20(String []i) {
    double p = Double.NaN;
    if (i[4] == null) {
      p = 0;
    } else if (i[4].equals("yes")) {
    p = WekaClassifier .N1ba01ea21(i);
    } else if (i[4].equals("no")) {
      p = 0;
    }
    return p;
  }
```

Listing 5-1 partial source code of j48tree generated in weka as java source code

Internally, Weka handles all values as doubles and it returns as double values. For instance from the above source code the Classifier () returns the values of P (predicted) as 0 and 1 when the source code is run in the java platform it gives the prediction value 0.0 or 1.0. Because it is declared as double value, here the problem is we don't know 0.0 or 1.0 is predicted to the correct class values with the attribute values. The class values and attributes values of the diabetes are nominal; which means all values are String data types. The solution for this problem is to create the Attribute value to pass it as an array of strings that lists the possible nominal values. The double that classification returns is the index of the chosen attribute in the original array.

If we had code that looked like this: String[] attributeValues = {"a", "b", "c"}; Attribute a = new Attribute ("attributeName", attributeValues); and classifyInstance () returned 2. Then the class it chose would be attributeValues [2] or c. With the distributionForInstance() method, the indexes of the two arrays match, so attributeValues [0] is the string name for the first element of the array returned or **a**.

```
public static double callingMethod(String sex, String age, String fbs, blurred_vision, String
family_history, String obesity, String ketone, String weight_loss) throws Exception{
switch (age) {
      case "child":
         Atrribute_val[1] = "child";
         break;
      case "young":
         Atrribute_val[1] = "young";
         break;
         case "adult":
         Atrribute_val[1] = "adult";
         break;
             case "old":
         Atrribute_val[1] = "old";
         break;
   }
```

Listing 5-2 Java code to convert double values into string

The above code shows that the way to mapping the double values of **classify(),** method while generating java source code to classify the diabetes types as Type1 or type2 based on the selected attributes values. The weka classifies as 0 or 1 which is the integer types. But this way of predicting may cause the problem how to identify the correct class type of diabetes. Therefore, it is necessary to convert the double values of the predicted class of diabetes types to its appropriate values (storing the array values as string data types).

5.1.1. Integration and Path setting

The most important task is the way to integrate the data mining results that is imbedded in java to prolog. There are so many ways to call the rules from weka through java neatbeanse, to integrate the rules generated by java or to access the predicted class values from weka. Through the following libraries like: Jasper, Java Log, Jinni(Java INference Engine and Networked Interactor), MINERVA, Prolog Café and JPL_call we can acces or call the predicted diabetes class types into prolog engine. From these libraries jasper and jpl_call is experimented to integrate the java code to prolog.

Jasper: jasper is a bi-directional interface between programs written in Java and programs written in Prolog. The Java-side of the interface constists of a Java package (jasper) containing classes representing the SICStus emulator. The Prolog part is designed as an extension to the foreign language interface, which means that foreign/3 declarations are used to map Java-methods on Prolog predicates and provide automatic conversion of arguments (Mixing Java and Prolog). However, this will most likely be extended in the future. The library does not yet define any predicates. However, it must be loaded to make sure that the JVM (the Java Virtual Machine) is loaded correctly and that the interface code has access to all relevant references to the JVM. It is loaded by the query.

|?- use_module(library(jasper)). This will dynamically load the JVM into memory and initialize the JNI(Java Native Interface) Invocation API.

JPL (Java Interface to Prolog) call: JPL is a library using the SWI-Prolog foreign interface and the Java jni interface providing a bidirectional interface between Java and Prolog that can be used to embed Prolog in Java as well as for embedding Java in Prolog. In both setups it provides a reentrant bidirectional interface (JPL 3.x Prolog API overview). Therefore jpl_call is selected for this study it itegrates java with prolog by the following method.

jpl_call (<Java Classname>, <method name>, [<String>], <Result>).

Path setting:

First of all setting of java, weka and prolog path by following the directory is very important in order to integrate java with weka and prolog with java as follows:

```
C:\ProgramFiles\Java\jdk1.8.0_45\bin;C:\ProgramFiles\Java\jre6;C:\ProgramFiles\Java\jre6\bin\client;C:\ProgramFiles\Java\jdk1.8.0_45\jre;C:\ProgramFiles\Java\jdk1.8.0_45\jre\bin\client\jvm.dll;
```

Listing 5-3 Java path setting

```
C:\Program Files\swipl\bin\jpl.dll;C:\Program Files\swipl\lib\jpl.jar;

C:\Research\all\WekaClassifier\wuy.pl;C:\Program Files\Weka-3-6\weka.jar;
```

Listing 5-4 prolog and weka path setting

```
C:\Research\all\WekaClassifier\src\wekaclassifier\WekaClassifier.java;

C:\Research\all\WekaClassifier\src\wekaclassifier\WekaClassifier.class;
```

Listing 5-5 path setting for integration

```
: - use module (library (jpl)). %this library is all we need to call Java from Prolog
Perform prediction (sex, age, fbs, blurred_vision, feamily_history, obesity, ketone, weight_loss ).
Perform prediction (sex, age, fbs, blurred_vision, family_history, obesity, ketone, weight_loss):-
jpl_call('jpl.Jpl',callingMethod,[sex, age, fbs, blurred_vision, family_history, obesity, ketone,
weight_loss],P),
  pred_value(P);
pred_value (P):-(P == 0.0) -> write('Your Prediction is type1 '),nl;
                (P == 1.0) -> write ('Your Prediction is type2 '),nl.
```

Listing 5-6 Calling predicted diabetes type from java to prolog

:- Use_module(library(jpl)). This library calls java from the prolog when the prolog consulting is running and added to the java path likes this: % Added [C:\Program Files\Java\jre1.8.0_60\bin\client,C:\Program Files\Java\jre1.8.0_60\bin] to %PATH% % library(jpl) compiled into jpl 0.06 sec, 1,959 clauses, after the path is added it returns True values this indicate that the prediction is ready to start.

Perform prediction (sex, age, fbs, blurred_vision, family_history, obesity, ketone, weight_loss). This is the facts to integrate with the calling method in the java calling method.

jpl_call('wekaclassifier.WekaClassifier',callingMethod,[sex, age, fbs, blurred_vision, family_history, obesity, ketone, weight_loss],P), is the library which call 'wekaclassifier. WekaClassifier' which embedded in the java, and call the predicted values of diabetes types returns P values as 0.0 or 1.0 and writes as type1 or Type2 based on the diagnosed criterions.

Knowledge Base: This is a container of rules about diabetes which are generated by J48 tree algorithm after calling via jpl_call library.

Knowledge Acquisition: This is the way the researcher Interviews with domain experts to acquire information about diabetes Management and Advisement as well as the way to categorize the attributes of diabetes used in chapter three from the expert knowledge.

Knowledge representation: This is the way knowledge of diabetes management and Advisement is represented from the decision tree after the diagnosing is predicted by data mining techniques (automatic rule generating by j48tree) then diabetes Advisement knowledge is represented from the decision tree.

Inference Engine: Is the way Prolog derive conclusions by backward chaining mode, starting from the facts to gets the conclusion that predicted by J48 classifier for the types of diabetes based on the attributes values asked by the prolog to the users via the user interface.

Explanation Facility: The explanation facility provides information to user for the questions asked by the prolog. The explanation is helpful to have clarity while answering to the questions asked by the system.

User Interface: The user interface can further be improved by adding a menu capability which gives the user a list of possible values for an attribute. It can further enforce that the user enter a value on the menu. This can be implemented with a new predicate, menu ask. It is similar to ask, but has an additional argument which contains a list of possible values for the attribute. It would be used in the program in an analogous fashion to ask: The user interface of a prototype is used as the means of communication between a user and a prototype integrated knowledge-based system. Its duty is to translate the Prolog rules of the prototype from its representation in the knowledge base and java back engine which may not be understandable by the end- users. The end users starts in Swi-prolog by typing “ask” then full stop and press enter then the system will display the welcoming window (discussed in the next chapter).

CHAPTER SIX

6. Implementation and Experimentation

Diabetes prediction and Advisement knowledge based system (DPATKBS), is capable of predicting diabetes as, Type1 and Type2. In addition, it provides Advisement and recommendation for the user about the result of predicted status of diabetes. The knowledge base contains validated rules and facts about diabetes Advisement.

The prediction process is undertaken by interacting with the user through presenting a serious of questions. The system asks the user by displaying question containing attributes with their values. The user is expected to reply for the questions or ask explanation for the questions.

Once diabetes types is predicted as Type1, or Type2 the system asks the user to continue the Advisement and get Advisement to manage diabetes.

The following figure shows the prediction process of the diabetes types and Advisement Advisements for the users.

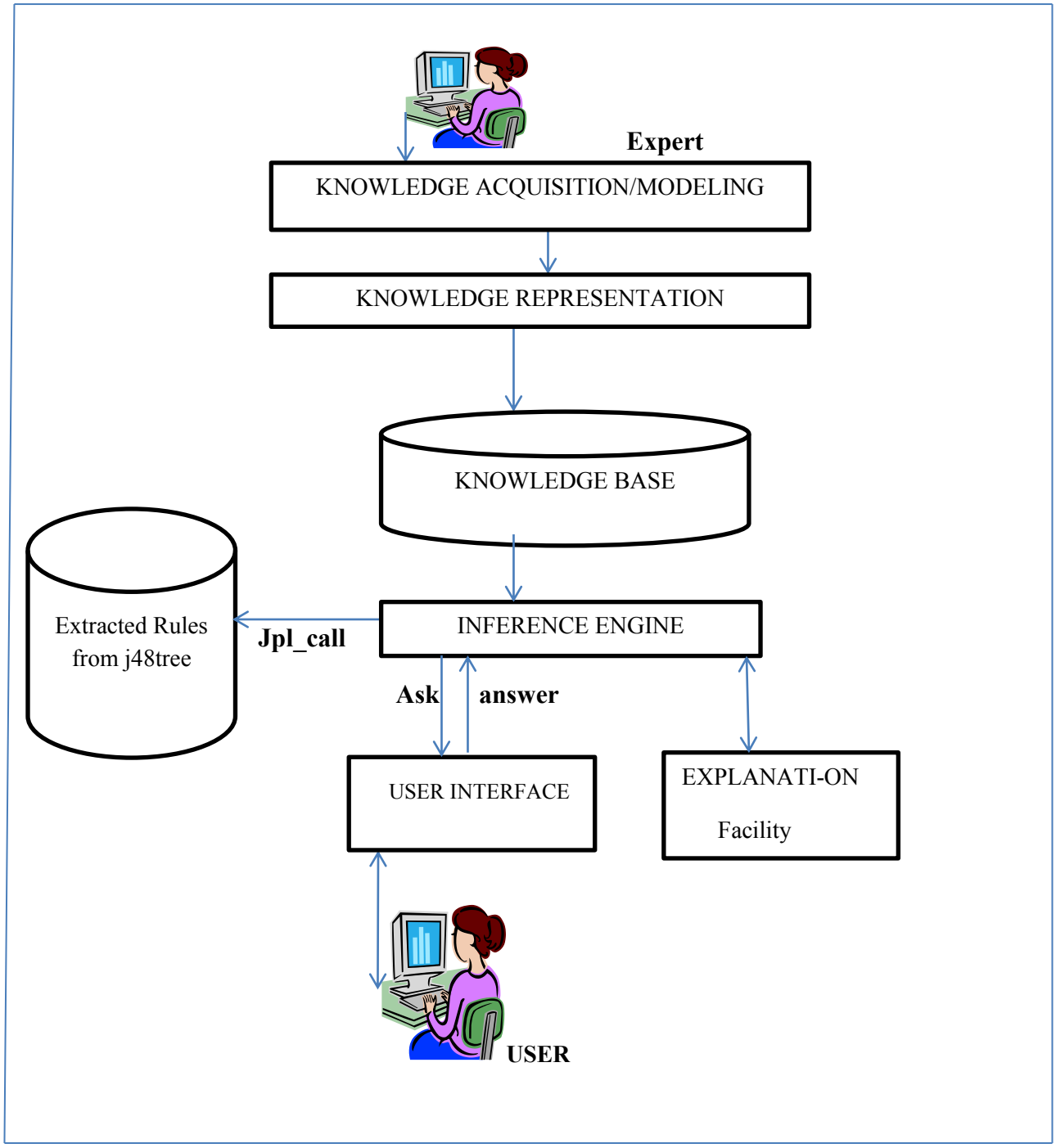

Figure 6-1 Architecture of DPATKBS (Diabetes Prediction and Advisement of knowledge base system)

The architecture of integrated data mining diabetes prediction with knowledge base system and Advisement is shown in figure 6.1 above. Knowledge acquisition, knowledge representation, knowledge base, inference engine, user interface and explanation facility are discussed below.

Extracted Rules: This are the rules that are generated from j48ree classifier while experiment is done on the data sets of diabetes .It is embedded in the java to call the predicted results via jpl_call library of swi-prolog.

Explanation Facility: This explanation facility provides information to user for the questions asked by the prolog. The explanation is helpful to have clarity while answering to the questions asked by the system. The user can get explanation of the question asked if he/she has no clarity with it. For example:

|do you want to know the reason you have type1 diabetes ?y/n /why

|: y.

Because your blood sugar is low, your age range is below 35, your ketone is existing and you have no blurred vision.

Figure 6-2 sample explanation facility

User Interface: As it is discussed in chapter five it is the interface between the users and the system in order to predict the types of diabetes after the user's type **ask** then type dot (.) then the system displays the following window to the end-users.

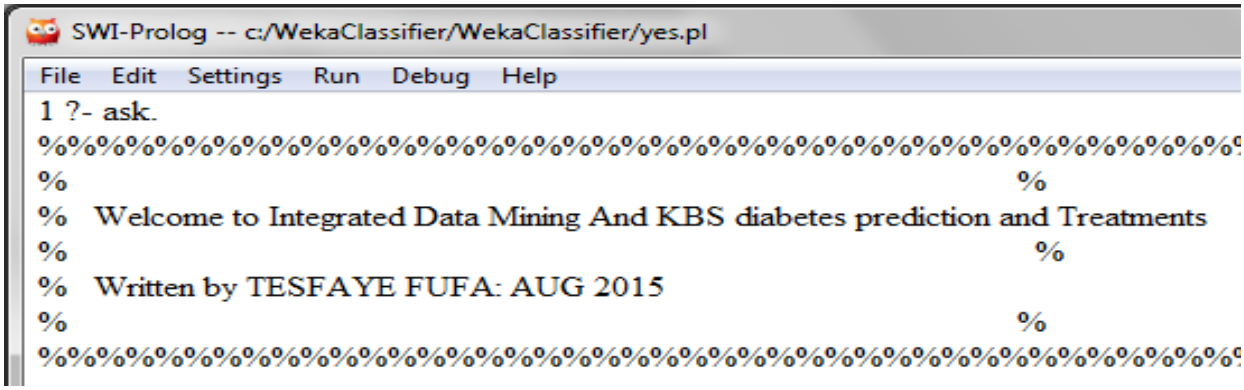

Figure 6-3 User interface of Diabetes Prediction and Advisement by integrating DM with KBS

After the user interface is displayed for the ende-user the system is ready to predict or diagnosing the diabetes types by asking different question after the user enter his/her name into the displayed window in order to continue the prediction process as follows:

Enter Your Name: Garamu.

WELCOME! Garamu Do you need Information about diabetes Prediction types? Y/n/
|: y.
Diabetes prediction is a way to predict the diabetes status as type1 or type2
Based on different attributes so that the Advisement system can use it.
Do you need to perform diabetes types Prediction or diagnose?! y/n
|: y.
Enter your sex: male/female
male->male,female->female
|: male.
enter your age range: child/young/adult/old
child->3-15,young->16-35,adult->36-55,old->greater than or equal 56
|: child.
Enter your range of fasting blood sugar: low/medium/high
low->below100 mg/dl, medium->greaterthan100 up to 300 mg/dl, high->greaterthan300 mg/dl
|: low.
say yes/no for your blurred_vision : yes/no
yes-> blurred, no->not blurred
|: no.
say yes/no for your family_history: yes/no
yes-> family_history_exist, no->family_history_notexist
|: no.
say yes/no for your obesity: yes/no
yes->obese, no->not_obese
|: no.
say yes/no for your ketone exists: yes/no

yes->exists, no->not_exist
|: yes.
say yes/no for your weight_loss exists: yes/no
yes->loss, no->not_loss
|: yes.
your prediction is **type1** diabetic do you want to continue Advisement of this diabetes type?/y/n/why

Figure 6-4 Dialog between the users and the system to predict the diabetes types

6.1.Testing and Evaluation of DPAT-KBS

In order to guarantee that DPAT-KBS meets the requirement it is developed for, it has to be tested. Testing is aimed at making sure amenability of the system with user expectation, human expert and system functioning (Hassan M. Ghaziri Elias M.Awad., 2013).It helps to answer the questions; is this the right system? And is the system right? For evaluating the performance of the knowledge base system test cases are prepared and given to the system. The outputs of the system are compared against domain area experts 'decision.

Table 6-1 Performance evaluation based on TP, FP, Precision, Recall and F-measure

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	Type
	0.923	0.065	0.923	0.904	0.914	type1
	0.935	0.077	0.935	0.949	0.942	type2
Weighted Avg.	0.93	0.072	0.931	0.93	0.931	

By Ignoring False Positive rate from the above table 6.2 the system registered 92.3% Recall and 92.3% TP rate as Type1 diabetes types, 92.3% **Precision** and 90.4% recall and 91.4% F-measure. The system also registered 93.5% TP, 93.5% **Precision**, 94.9% Recall and 94.2 % F-measure as type2 diabetes. To summarize the performance of the system based on different criterion, the researcher take the precision and TP rate as better accuracy rather than other methods of evaluations with the accuracy of 93.5% as type2 and 92.3% as type1 diabetes.

User Acceptance Testing (UAT)

User acceptance testing (UAT) is the last phase of the software testing process**,** the aim of undertaking user acceptance testing is to make sure how well DPAT-KBS is performing on the eyes of users so as to make sure that the system is accepted and usable by users. Five domain experts are selected to test the system. The evaluators assessed DPAT-KBS by using the following standards.

1. Simplicity to use and interact with the system.
2. Attractiveness of the system
3. Efficiency in time
4. The accuracy of the system in reaching a decision to identify the types of Diabetes
5. Inclusion of suggestion and important Advisement about diabetes types.
6. The ability of the system to make right conclusion and recommendation
7. Importance of the KBS in the domain area

Evaluation of the system using user acceptance testing is done to make sure how the users or domain experts 'view the system on the base of the above-mentioned evaluation standards. Different researchers have used different types of user acceptance testing evaluation criteria. But for this study, the evaluation criteria suggested by Solomon (Solomon Gebremariam, 2013) Pu et al. (Chen, 2010) and Rediet (Rediet A., 2006) is customized and used to ease the evaluation process. The weights scale Suggested by Solomon (Solomon Gebremariam, 2013) has been used such that Excellent = 5, Very Good =4, Good =3, Fair =2 and Poor =1 Table 6.3 depicts that summary of domain experts'evaluation of the system. The values indicate numbers of evaluator who evaluated the system as poor, fair, good, very good and excellent with respect to evaluation criteria.

Table 6-2 summary of domain experts 'evaluation of the system

No-	Criteria of evaluation	Poor (1)	Fair (2)	Good (3)	Very good (4)	Excellent(5)	Average
1	Simplicity to use and interact with the system.	1	2	3	0	0	2.8
2	Attractiveness of the system	0	2	2	2	0	3.6

3	Efficiency in time	0	0	0	0	5	5
4	The accuracy of the system in reaching a decision to identify the types of Diabetes	1	2	1	3	2	6
5	Inclusionof suggestion and important Advisement about diabetes types.	0	0	2	2	1	3.4
6	The ability of the system to make right conclusion and recommendation	0	0	1	3	2	5
7	Importance of the KBS in the domain area	0	0	1	3	2	5
						Total Average	4.5

6.2. Discussion

Therefore from table 6.3, 20% of evaluators replied poor for **simplicity to use and interact with the system**. The same percentage, 40%, of respondents replied Fair and Good for it. This is due to the command line interface used for interaction with the user such that users are expected to type commands and replies. With regard to the second question, that is **Attractiveness of the system** 40%, 20%, 40% of evaluators replied fair, good and Very Good. The third question is about **Efficiency of the system in terms of time**. All of the evaluators (100%) agreed that, the efficiency of the system is excellent in replying for their request.

The fourth evaluation criteria is about **the accuracy of the system in reaching a decision to identify the types of diabetes** and 60% of evaluator scored very good and 40% of them as excellent. Whereas 60%, which is the majority, of the evaluator scored the system as very good and 20% as good and excellent for **inclusion of suggestion important Advisement about diabetes prediction types.** The other evaluation criterion is **the ability of the system to make the right conclusions and recommendation.** Among the evaluators**,** 60% replied very well for PDAT-KBS and 40% of them replied excellent for KBS.

The last criterion is i**mportance of the KBS in the domain area.** This criterion is included to measure how PDAT-KBS is important in the area of diabetes prediction. And 60% of evaluators replied very good and 40% replied good. This implies that the contribution of developing KBS like PDAT-KBS is important in the area of diabetes prediction and advising of users after predicting diabetes as a type1or type2. Finally, according to the evaluation filled by domain experts', PDAT-KBS has registered 4.5 out of 5(90%), which is taken as very good achievement.

6.3.Challenges and Limitations

In this study promising results are achieved in integrating machine learning convinced patterns with knowledge based system for predicting diabetes types and providing advisement to domain experts. Some challenges have been encountered which hinder the system from scoring a better achievement. The first one is in the course of integration, two interfaces namely; graphical user interfaces (for the integrator java) and command line interface for PDAT-KBS has been used. A challenge has been encountered in bringing the integrator (java) and PDAT-KBS together under one interface. This is reflected on user acceptance test in that evaluator rated the simplicity to use and interact with the system below very good.

CHAPTER SEVEN
CONCLUSION AND RECCOMENDATION

7. Conclusion

In this study, the possibility of integrating data mining models with knowledge based system is explored. For the integration process samples of Diabetes dataset was taken from Addis Ababa Black Lion Diabetes Clinic Center. The dataset is preprocessed and made suitable for mining steps. Due to several limitations in gaining knowledge for knowledge base from domain experts in the area of diabetes prediction, an automatic knowledge acquisition mechanism is proposed in this study. Data mining has proven to induce hidden knowledge from large collection of dataset. Hence, data mining classifier, J48 is employed for knowledge acquisition step since it has performed best among the selected classifiers with an accuracy of 93.47%. Rules generated and embedded in java application have enabled building diabetes prediction and advising knowledge based system through swi-prolog. Following the successful integration of induced knowledge with the knowledge based system, the rules based from j48 tree diabetes prediction and advising KBS is built. System performance testing is undertaken to make sure that the right PDAT-KBS has been built. Hence the testing disclosed that the system has 90% of accuracy with very good Precision and Recall. 93.2% User acceptance testing is achieved based on seven criteria of evaluation. Selected domain experts are trained and used the system to evaluate how much the KBS meets their requirements.

The system has registered 90.16% overall accuracy according to the system performance and user acceptance tests. The performance analysis shows that DPAT-KBS registered acceptable performance. The knowledge base system also provides Advisement and information based on the new changes yielding upto-date knowledge. This characteristic distinguishes this study in that it somehow alleviated problem of diabetes prediction which are claimed to have limitation in predicting new types of diabetes. However, further exploration and study has to be done to refine and yield a better knowledge based system which can be deployed in real diabetes medical centers and provide Advisement to experts and patients, so that they can take timely and appropriate actions.

7.1. Recommendation

The designed prototype KBS supports two types of diabetes namely; Type1 and Type2. Hence the researcher believes further researches have to be done to improvement the benefits of integration of data mining with knowledge based system and the following are recommended for further study.

- Building hybrid knowledge based system which is capable of employing rule based reasoning and case based reasoning with integrated data mining techniques.
- Building KBS with graphical user interface which is simple to use and attractive to users using prolog.net.
- The rules in the knowledge base of the prototype system should be updated with the aim of learning from experience so as to update its knowledge. Therefore, additional research should be accomplished to update the rules of the system automatically.
- Making the system to access from anywhere and anytime where the connection is available is better. Therefore developing web-based diabetes diagnosis and Advisement as future work is appropriate.

References

A. Aamodt, E. Plaza. (1994). (Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications. IOS Press, 7*, 39-59.

A. Ranjan, J. G. (2002). Knowledge Acquisition, Representation and Reasoning.

A.N. Masizana-Katongo, T. L.-D. (2009). An Expert System for HIV and AIDS Information. *Proceedings of the World Congress on Engineering,*, *1*, pp. 1-5.

Agrawal, R. &. (1994). Algorithms for mining association rules. . *Proceedings of the 20th VLDB Conference:* , (pp. 1-13.).

Agrawal, R. I. (1993). Mining association rules between sets of items in large databases. . (pp. 1-10.). Proceedings of the 1993 ACM SIGMOD conference:.

ahmed, D. (2010). Knowledge based Diagnosis of Abdomen Pain using Fuzzy Prolog Rules. *Journal of Emerging Trends in Computing and Information Sciences.*

Akerkar, Eds. Sajja. (2010.). Knowledge-Based Systems for Development Systems. In K.-B. S. Knowledge.

Akerkar, P. (2010). *Knowledge-Based Systems for Development.*

Akerkar, P. Heijst. (2010). Knowledge-Based Systems for Development.

Akerkar, P. S. (2010.). Advanced Knowledge Based Systems: Model, Applications & Research. *1*, pp. 1 – 11,.

al Brown, S. A. (2013). Diabetes in Ethiopia: overcoming the problems of care delivery.

Alechina, N. (2012). *Knowledge Acquisition, Representation, and Reasoning.*. united kingdom.

Aref, R. (2000). A Knowledge based system for comfort analysis of internal environment of hotels.

ATLAS, I. D. (2013). International Diabetes Federation.

atzilygeroudis. (2007). Integrations of Rule-Based and Case-Based Reasoning. *Greece Research Academic Computer Technology Institute.*

AZevedo&Santos. (2008). *KDD, SEMMA AND CRISP-DM: A Parallel Overview.*

Baran Hashemi, H. J. (2012). An Approach for Recommendations in Self Management of Diabetes based on Expert System. *International Journal of Computer Applications, 53*.

Bernander, O. (2009). Microsoft Encarta 2009: neural network. Microsoft Corporation.

Berry and Linoff. (2004). *Data Mining Techniques for Marketing, Sales and CRM, 2nd Ed.*

Bethesda. (2013.). *www.diabetes.niddk.nih.gov. [Online] National Diabetes Information Clearinghouse*. Retrieved from www.diabetes.niddk.nih.gov

Boose, B. G. (1988). *Knowledge Acquisition for Knowledge-Based Systems* (Vol. 1). London, new york: Academic Press.

Bramer, M. (2007). Principles of data mining.

Chen, P. a. (2010). "A User-Centeric Evaluation Framework of Recommender Systems,". *In Proceedings of the ACM RecSys 2010 workshop on User-centeric Evaluation of Recommender Systems and Their Interfaces (UCERSTI)*, pp. 157-164.

Chen, S. (2007). Knowledge Representation and Reasoning Methodology based on CBR Algorithm for Modular Fixture Design. *Journal of the Chinese Society of Mechanical Engineers*, 593-604.

Corporation, T. C. (1999). *Knowledg Discovery, Introduction to Data Mining and Knowledge based system.*

Daca. (2004). Standard Treatment Guidlines for District Hospital in Ethiopia.

Danubianu, M. (2008). Design of An Expert System for Efficient selection of Data Mining Method.

David Hand, H. M. (2001). *Principles of Data Mining.* MIT Pres.

Dereje Abebe, Y. T. (2005). Diabetes Mellitus For the Ethiopian Health Center Team. *In collaboration with the Ethiopia Public Health Training Initiative, The Carter Center*. Debub University: Funded under USAID Cooperative Agreement No. 663-A-00-00-0358-00.

Dr. Sudhir B. Jagtap. (2013). Census Data Mining and Data Analysis using WEKA. *International Conference in "Emerging Trends in Science, Technology and Management*.

E, S. (2011). "Prototype of Knowledge Based System for Axiety Mental Disorder Diagnosis, School of Informatin Science, Addis Ababa University, Addis Ababa, Msc Thesis.

E. Turban. (1990). "Decision Support and Expert Systems: Management Support Systems",.

E.A. Feigenbaum and R.S. Engelmore. (1993). *Expert System and Artificial Intelligence.* japanese Technology Evaluation Center panel‟s report about Knowledge-Based Systems.

Fogel, D. B. (2006). Evolutionary Computation. In *Defining Artificial Intlligence.* The Institute of Electrical and Electronics Engineers,.

Gau, M. (1990). Knowledge Acquisition for a Diagnosis-Based Task.

Getachew1, S. M. (2014.). High magnitude of diabetes mellitus among Active pulmonary TB patiants in Ethiopia.

Gilbar, B. (1999). Flexible Knowledge Acquisition Through Explicit Representation of Knowledge. *Roles. Marina del Rey: USC/Information Sciences Institute.*

Gilman, D. M. (2000). Data Mining Overview.

Glance, Africa at. (2013.). Diabetes prevalence , second edition.

Gomez, A. (1994). From Knowledge Based Systems to Knowledge Sharing Technology. *Evaluation and Assessment.*

Govindan, V. (2012). Using Rule Based Reasoning and Object Oriented. *Journal of Social Sciences 8 (1)*, 66-73.

Han, J & Kamber, M. (2006). *Data mining: concepts and techniques.* San Francisco:.

Han, J & Kamber, M. . (2006). Data mining: concepts and techniques. (2nd ed.). *San Francisco:Morgan Kaufmann Publishers*.

Hand, D. M. (2001). *Principles of data mining.* Massachusetts Institute of Technology press.

Hassan M. Ghaziri Elias M.Awad. (2013, Seb 17 THU). Retrieved Seb 17, 2015, from http://books.google.com.et/books?id=CI63F2n4N7AC&pg=PA264&lpg=PA264&dq=

Henok, B., & Burge. (2011, 1998). A case based reasoning knowledge based system for hypertension.

Ibrahim M. Ahmed, A. M. (2014). Development of an Expert System for Diabetic Type-2 Diet. *International Journal of Computer Applications (0975 – 8887)*.

Int. J. Man. (1987). a knowledge-acquisition tool for expert systems. 12.

Jackson, J. (1999). *DatAa Mining Conceptual Overview.*

JPL 3.x Prolog API overview. (n.d.). Retrieved september 10, 2015, from http://www.swi-prolog.org/packages/jpl/prolog_api/overview.html#jpl_call4

K. Rajesh, V. Sangeetha. (2012). Application of Data Mining Methods and Techniques for Diabetes Diagnosis. *International Journal of Engineering and Innovative Technology (IJEIT), 2*(3).

Kalayou Kidanu Berhe, Alemayehu Bayeray Kahsay. (2013). Adherence to Diabetes Self-Management Practices in ethiopia. *Greener Journal of Medical Sciences,*.

Kalayou Kidanu, B. A. (2013). Adherence to Diabetes Self-Management Practices in ethiopia. *Greener Journal of Medical Sciences,*.

Kantardzic, M. (2003). Data Mining: concepts, models, methods, and algorithms. In W. &. Sons.

Kohavi, R. (2001). Data mining for knowledge based system.

Kotsiantis, S. &. (2006). Association rules mining: a recent overview. *GESTS International Transactions on Computer Science and Engineering*, 71-82.

Kowalski, R. (1988). The Early Years of Logic Programming", Communications of the ACM. *,Vol. 31, No. 1, pp. 38-43,*.

Kurgan, L. A. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*.

Larose, Daniel T. (2005). Discovering knowledge in data: an introduction to data mining. *New Jersey: John Wiley & Sons.*

Lech, M. (2000). Validation of rile based system generated by classification algorithms.

M. Durairaj, V. Ranjani. (2013, OCTOBER). Data Mining Applications In Healthcare Sector: A. *International Journal of Scientific & Technology Research, 2*(10,).

Mak, M. (2010). Applications of Knowledge-Based Expert Systems to Feng Shui Knowledge.

McCarthy, J. (1960). Recursive Functions of Symbolic Expressions and Their Computation by Machine. *Communications of the ACM, 3*, pp. 184–195, 1960.

Megerssa et al., J. r. (2013.). Prevalence of Undiagnosed Diabetes Mellitus and its Risk Facto.

Mixing Java and Prolog. (n.d.). Retrieved Sept 2, 2015, from https://sicstus.sics.se/sicstus/docs/3.7.1/html/sicstus_12.html

Naheed G., J. A. (2010). Knowledge, Attitudes and Practices of Type 2 Diabetics Mellitus,.

O, M. Mehdi Owrang. (2008). Database systems techniques and tools in automatic knowledge acquisition for rule-based expert system,in Knowledge-Based Systems Vol 1. *1*, 201-248.

Osuagwu, E. (2006). The Underlying Issues in Knowledge Elicitation. *Interdisciplinary Journal of Information, Knowledge, and Management.*

Port Louis, Mauritius. (2009). International Diabetes Conference & Assocaited Diseases.

Prasanna Desikan, K.-W. H. (2011). Data Mining for Helathcare Management.

Prasanna Desikan, Kuo-Wei Hsu,Jaideep Srivastava. (2011). Data Mining for Healthcare Management.

R. Jayabrabu, D. V. (2012). A Framework: Cluster Detection Cluster Detection and Multidimensional Visualization of Automated Data Mining Using Intelligent Agents. *International Journal of Artificial Intelligence & Applications (IJAIA)*.

R.A. Akerkar, P. S. (2010). Knowledge-Based Systems. *Jones and Bartlett Publishers*.

Raf.R. (2012). *Mobile Data Mining for Intelligent Healthcare.*

Ranjan.J. (2006). Knowledge Acquisition, Representation, and Reasoning.

Rediet A. (2006). Design and Development of a Prototype Knowledge-Based System for HIV Pre Test Counseling,School of Information Science, Addis Ababa University, Addis Abab, Msc Thesis 2006.

Refaat, M. (2007). Data preparation for data mining using SAS. San Francisco. *Morgan Kaufmann Publishers.*

Reganie, B. (2013). Application of Data Mining Techniques for Customers Segmentation and Prediction the case of Buusaa Gonofa Microfinance Institution.

Reganie, B. (2013). Application Of Data Mining Techniques For Customers Segmentation And Prediction The Case Of Buusaa Gonofa Microfinance Institution.

Richard et, a. (2010). Methods of Conceptual Modeling.

S. Robertson and J. Kingston. (2012). *"Selecting a KBS Tool Using a Knowledge-based System",*. Retrieved from Available at: http://www.aiai.ed.ac.uk/project/ftp/documents/1997/97-paces97-select-kbs-tool.ps,.

Sajja, R. A. (2010). "Knowledge-Based Systems". *ones and Bartlett Publishers,*.

Salem, A. (2007). Case Based Reasoning Technology for Medical Diagnosis. . *World Academy of Science, Engineering and Technology.*

Sandoval, L. (1995). Knowledge Acquisition,Representation,and Reasoning. a graduate student at California State University, Long Beach.

Saxil. (2010). Building blocks of knowledge based system architecture.

Schreiber, P.-H. S. (2011). Conceptual Modelling for Knowledge-Based. Unilever Research Vlaardingen, University of Amsterdarm,CIBIT, Bolesian.

Seblewongel, E. (2011). *prototype knowledge based system for anxiety mental disorder diagnosis.* Ethiopia, Addis Ababa: Addis Ababa University.

ShusakuTsumoto and Shoji Hirano. (2011). Temporal Data Mining in Hospital Information Systems.

Simon, S. (2004). Case-Based Reasoning: Concepts, Features and Soft Computing. . ,pp 233–238.

Solomon Gebremariam. (2013). Self-Learning Knowledge Based System for Diagnosis and Treatment of Diabets, School of Information Science, Addis Ababa University,M Sc Thesis 2013.

Sonmenz, C. (1988). Artificial Intelegence Notes Reasoning Methods.

Tan, H. C. (2005). Data Mining Applications In health care.

Tehrani, S. (2009). A Conceptual Model of Knowledge Elicitation. . *College of Business, Technology and Communication, .*

Thirumal P. C. and Nagarajan N. (2015). Utilization of data mining techniques for diagnosis of diabetes mellitus - a case study. *ARPN Journal of Engineering and Applied Sciences, 10*(1).

Tu Bao Ho, S. K. (2007). *Knowledge Acquisition by Machine Learning and Data Mining.*

U. Nilsson and J. Maluszynsk.i. (2000). *Logic, Programming and Prolog", 2nd ed, John Wiley &Sons, Sweden, .*

Velide, V. K. (2014). A Data Mining Approach for Prediction and Treatment of diabetes Disease. *IJSIT*.

Wang, S. (2011). Knowledge elicitation approach in enhancing tacit knowledge sharing. *Emerald,*, pp. 1039-1064.

Witten, I. H. & Frank, E. (2005). Data mining: practical machine learning tools and techniques.(2nd ed.). *San Francisco: Morgan Kaufmann Publishers.*

APPENDIX I

Variables and descriptions of dataset

Attribute number	ATTRIBUTES	DESCRIPTION	TYPE	UNIT
1	Sex	Patient's sex	Nominal	
2	AGE	Age of patient(years)	Numeric	-
3	FBS	Diabetes indicator, glucose level in blood(Dependent variable) or 8-hour Fasting plasma glucose,	Numeric	Mg/dl
4	FAMILY_HISTORY	Genetic(existence of diabetes with family)	Nominal	-
5	OBESITY	Overweight or fatness	Nominal	-
6	KETONE	Byproduct produced when the body uses large amounts of fat as fuel.	Nominal	-
7	WEIGHT_LOSS	Whether the patient reduced in weight or not	Nominal	-
8	Blurred-vision	Whether or not the patient have eye problem	Nominal	-
9	Class	Categories of Diabetes types (type1,type2)	Nominal	-

APPENDIX II

Detail Advisement for type1 and type2 diabetes with medication and non-medication methods, for both non-medications Advisement is the same, only the difference is on exercise and medication Advisement.

Diabetes Type	Advisement and management methods
Type1	**A. Non-Drug-Therapy (**Medication**) Advisement.** **Diet Advisement**: Diet is a basic part of management in every case. Advisement cannot be effective unless adequate attention is given to ensuring appropriate nutrition**.** **Food items the diabetic should avoid** (rapidly absorbed carbohydrates) Sugar, honey, jams, candy, marmalade Cakes, Sweet Biscuits Soft drinks (Coca Cola, Miranda etc) Alcohols **Foods which the diabetic can take with restrictions Food items from grains**: enjera, bread, kinche, kita, atmit Foods items prepared from peas, beans, lentils, chick peas Potato, sweet potato, kocho, bulla . **Food items the diabetic can take freely or with minimal restrictions** : Lean meat and fish Eggs, milk, cottage cheese Green leafy vegetables (cabbage, tomato, pumpkin, carrots, onion) Tea, coffee and lemon juice without sugar, Ambo water, other mineral waters Spices: pepper, garlic, 'berbere'. **Education:** Education of the person with diabetes is an essential component of management in every case. To ensure appropriate management, the basic knowledge and skills should be acquired by the patient and his family and the health care team should work closely with the patient to achieve this objective and to promote self-care. The person with diabetes should also be involved in setting therapeutic targets for weight, blood pressure and blood sugar control. **B. Drug-Therapy(Medications) Advisement:** **Insulin therapy:** In type 1 diabetes, the body loses the ability to produce

	insulin, thus, exogenous insulin must be administered indefinitely. A standard insulin Advisement consists of one or two injection/day, using intermediate or long acting insulin with or without regular insulin. In type 2 diabetes, insulin may be necessary on a long term basis to control glucose levels if diet and oral agents have failed. In addition, some patients whose type 2 diabetes is usually controlled by diet alone or diet and an oral agent may require insulin temporarily during illness, infection, pregnancy, surgery or some other stressful events . Type 1 DM Patients with type 1 diabetes have on absolute requirement of insulin for survival. Insulin is also used in type 2 diabetics when a combination of oral agents fails to achieve glucose targets and temporarily in patients with serious infection or surgery. The types of insulin available are rapid acting, short acting, intermediate acting and long acting. Standard insulin therapy consists of one to two injections per day using intermediate or long acting insulin with or without regular insulin. Starting insulin does vary from 0.15 to 0.50/kg (as high as 1.5 u/kg in cases of severe insulin resistance) depending on patient size and degree of glycaemia. Adults of normal weight may be started with 20-25 u/d of intermediate acting insulin and increased to maintain a blood sugar level of 80-120 mg/dl. **Exercise:** Physical activity promotes weight reduction and improves insulin sensitivity, thus lowering blood glucose levels. Together with dietary Advisement, a programme of regular physical activity and exercise should be considered for each person. individuals with type 1 diabetes should monitor blood glucose before, during, and after exercise, delay exercise if blood glucose is > 250 mg/dL, <100 mg/d), or if ketones are present, eat a meal 1 to 3 hours before exercise and take supplemental carbohydrate feedings at least every 30 min during vigorous or prolonged exercise decrease insulin doses (based on previous experience) before exercise and inject insulin into a non-exercising area..

Type2	**Type 2 DM** : Provided pharmacologic therapy is not required immediately all patients should be given at least a one month trial of diet, exercise and weight management. If this regimen does not lead to adequate blood glucose control, oral antihyperglycemic agents with or without insulin are indicated. Insulin may be needed in symptomatic patients who have type 2 DM with FBG values greater than 250 mg/dl. The common ant hyperglycemic agents in use are discussed below. **Blood sugar monitoring:** There are two common techniques used by doctors to control the sugar level in the blood of individuals living with diabetes. These are frequent measurements of blood glucose and measurement of hemoglobin A1C **.** **Frequent measurements of blood glucose:** The amount of sugar in the blood can be measured before fasting. The patient should not eat food for at least 8 hours before taking sample blood for testing. The aim is to make the fasting sugar level in the blood between 70 and 120 mg/dL. **Measurements of hemoglobin A1C:** It is used to measure the average sugars level that was accumulated in our blood for the last 2-3 months by looking at the level of sugar in our hemoglobin. Hemoglobin is a protein that is found in red blood cells and its function is to transport oxygen within the body. In the normal case, the level of glucose in the blood is low, but, if it is abundant it will stick with the hemoglobin. Hemoglobin A1C test tells the amount of glucose that had stacked with the hemoglobin within the last 2-3 months.

APPENDIX III

Interview questions: After introducing the objective of the study, respondents were requested to participate by responding the interview questions. The answers of the respondents were recorded by using paper and pen for the following interview questions. The following interview questions are the main area to cover how general practitioners diagnose a patient.

1. What is Diabetes?
2. What are the major causes for Diabetes **Disease?**
3. How can we differentiate Type 1 with Type 2 diabetes during diagnosis?
4. What are the main challenges encountered during diagnosis of diabetes?
5. What are the main techniques to control diabetes type1?
6. What are the main techniques to control diabetes type2?

APPENDIX IV

Questionnaire to test and validate the performance of the Integrated Data mining and knowledge based system to predict and Advisement of diabetes with rule based knowledge based system

1. Is the system easy for you to interact with it?

1. Poor 2. Fair 3. Good 4. Very good 5. Excellent

2. Is DPAT-KBS attractive?

1. Poor2. Fair 3. Good 4. Very good 5. Excellent

3. Is the system more efficient in running time? 1. Poor 2. Fair 3. Good 4. Very good 5. Excellent

4. Does the system incorporate sufficient knowledge to categorize a given diseases into subcategory that have a prototype? 1. Poor 2. Fair 3. Good 4. Very good 5. Excellent

5. Is the system provides the right description and Advisement patient need to follow while diagnosis by human expert. 1. Poor 2. Fair 3. Good 4. Very good 5. Excellent

6. How do you rate the significance of the system in the domain area? 1. Poor 2. Fair 3. Good 4. Very good 5. Excellent

7. How is DPAT-KBS different from a diagnosis conducted by human expert?

8. What issues are covered by the advisor service of the system?

9. Do you feel any uncovered areas by the prototype about patient diagnosis system? If you feel please state them.

10. Does the system are have any significance in the domain area?

12. What is the strength of DPAT-KBS?

13. What are the limitations of DPAT-KBS?

APPENDIX VI

Partial integration code

```
:-ensure_loaded(library('jpl.dll')).% library which loads VM(virtual machine)

:-ensure_loaded(library(jpl)).% library which loads integrated rules through java interface to swi-prolog

read_name(Name),

write('WELCOME! '), write(Name),

diabetes_question,

read_int(Int),

check_interest1(Int, Action1),

Action1,

read_ans(Ans),

check_interest2(Ans, Action2),

Action2,nl,

    %Advisement_question,

    %read_ans(Ans),

   % check_interest3(Ans, Action3),

    %Action3,

    nl.

%Question Section

diabetes_question:- write(' Do you need Information about diabetes Prediction types! y/n'),nl.

prediction_question:- write(' Do you need to perform diabetes types Prediction! y/n'),nl.


%Information Section
```

```
Advisement_info:-

write('diabetes prediction is a way to predict the diabetes status as type1 or type2'),nl,

write( 'based on different atrributes so that the Advisement system can use it.'),nl,prediction_question.

prediction_info:-write('Prediction Started'),nl;

predictor.

%Input Section

read_name(Name):-  write(' Enter Your Name: '), read(Name),nl.

read_int(Int):- read(Int),nl.

read_ans(Ans):-read(Ans),nl.

%Checking Section

check_interest1(Int, Advisement_info ):- Int = 'y'.

check_interest1(Int, prediction_question):- Int = 'n'.

check_interest1(_,"").

check_interest2(Int, predictor ):- Int = 'y'.

check_interest2(Int, end):- Int = 'n'.

check_interest2(_,"").


%Prediction Section

predictor:-

write('Enter your sex: male/female'),nl,

write('male->male,female->female'),nl,

read(sex),nl;

write('enter your age range: child/young/adult/old'),nl,
```

```
write('child->3-15,young->15-40,adult->40-59,old->greater than or equal to 60'),nl,

read(age),nl;

write('Enter your range of fasting blood sugar: low/medium/high'),nl,

write('low->below100, medium->greaterthan100or200, high->greaterthan200'),nl,

read(fbs);

write('say yes/no for your blurred_vision : yes/no'),nl,

write('yes-> blurred, no->not blurred'),nl,

read(blurred_vision),nl;

write('say yes/no for your family_history: yes/no'),nl,

write('yes-> family_history_exist, no->family_history_notexist'),nl,

read(family_history),nl;

write('say yes/no for your obesity: yes/no'),nl,

write('yes->obese, no->not_obese'),nl,

read(obesity),nl;

write('say yes/no for your ketone exists: yes/no'),nl,

write('yes->exists, no->not_exist'),nl,

read(ketone),nl;

write('say yes/no for your weight_loss exists: yes/no'),nl,

write('yes->loss, no->not loss'),nl,

%read(weight_loss),nl;

%write('your prediction is type1 diabetic because P values=:0.0').

read(pred_value),nl;

perform_prediction(sex, age, fbs, blurred_vision, family_history, obesity, ketone, weight_loss ).
```

```
perform_prediction(sex, age, fbs, blurred_vision, family_history, obesity, ketone, weight_loss):-

jpl_call('yes.Yes',callingMethod,['sex, age, fbs, blurred_vision, family_history, obesity, ketone, weight_loss'],P),

write('Works Here'),nl,nl,

pred_value(P),

pred_value(P),Advisement_question.

pred_value(P):-

(P == 0.0) -> write('Your prediction is type1 DM'),nl,Advisement_question1;

(P == 1.0) -> write('Your prediction is type2 DM'),nl,Advisement_question2.

Advisement_question1:-write('Do you need Advisement suggestion on the current diabetes prediction status! (y/n/why):'),nl,

read(Advisement_question),

((Advisement_question=='y',afterpred_value1);

(Advisement_question=='n',write('Good bye'),nl,end);

(Advisement_question=='why',write('Advisement suggestion is a way to provide Advisement Advisement for current prediction,'),nl,

write('If you do not want to see the Advisement suggestion by only saying n->not'),nl,

write(' you can abort here the system or you are interesting only on prediction, so'),nl,why)).

why:-write('Do you need Advisement suggestion on the current diabetes type status! (y/n):'),nl,

read(Advisement_question1),

((Advisement_question1=='y',afterpred_value1),nl;

(Advisement_question1=='n', write('Good bye'),nl,end)).

Advisement_question2:-write('Do you need Advisement suggestion on the current diabetes status! (y/n/why):'),nl;

read(Advisement_question2),nl;
```

((Advisement_question2=='y',afterpred_value2);

(Advisement_question2=='n',write('Good bye'),nl,end);

Advisement_question2:-write('Do start need management on the current diabetes prediction status! (y/n/why):'),nl,

read(Advisement_question2),

((Advisement_question2=='y',afterpred_value2);

(Advisement_question2=='n',write('Good bye'),nl,end);

(Advisement_question2=='why',write('Advisement suggestion is a way to provide management Advisement for current prediction of diabetes,'),nl,

write('If you do not want to see the Advisement suggestion by only saying n->not'),nl,

write(' you can abort here the system or you are interesting only on prediction, so'),nl,why1)),nl.

why:-write('Do you need Advisement suggestion on the current diabetes types status! (y/n):'),nl,

Advisement_question1:write('========Recommendation on Non_Drug Theraphy======='),nl,nl,

write('FOR DIET Advisement:'),nl,

write('Food items the diabetic should avoid :'),nl,nl,

write('rapidly absorbed carbohydrates'),nl,nl,

write('Sugar, honey, jams, candy, marmalade'),nl,nl,

write('Cakes, Sweet Biscuits Soft drinks (Coca Cola, Miranda etc) ¬ Alcohols'),nl,nl,

write('Foods which the diabetic can take with restrictions ¬ Food items from grains'),nl,nl,

write('enjera, bread, kinche, kita, atmit ¬ Foods items prepared from peas, beans, '),nl,nl,

write('lentils, chick peas ¬ Prediction otato, sweet potato, kocho, bulla'),nl,nl,

write('Food items the diabetic can take freely or with minimal restrictions:'),nl,nl,

write('Lean meat and fish ¬ Eggs, milk, cottage cheese'),nl,nl,

write('Green leafy vegetables (cabbage, tomato, pumpkin, carrots, onion) '),nl,nl,

write('Tea, coffee and lemon juice without sugar'),nl,nl,

write('Ambo water, other mineral waters ¬ Spices: pepper,

garlic, ‘berbere’'),nl,nl,

write('==================Recommendation on education========'),nl,nl,

write('By the diabetes patients should make a team and discuss eachother or education should be given by domain experts at Advisement time'),nl,nl,

write(''),nl,

write('===========Recommendation On Exercise======='),nl,

write('Exercise To avoid exercise-related hyper- or hypoglycemia, '),nl,

write('individuals with type 1 diabetes should monitor blood glucose before, during, and after exercise, delay exercise if blood glucose is > 250 mg/dL, <100 mg/d), or if ketones are present'),nl,

write('eat a meal 1 to 3 hours before exercise and take supplemental carbohydrate feedings at least every 30 min during vigorous or prolonged exercise decrease insulin doses '),nl,

write('==========================Recommendation on Drug_Theraphy/Insulin============='),nl,

write('A standard insulin Advisement consists of one or two injection/day'),nl,

write('Starting insulin does vary from 0.15 to 0.50/kg (as high as 1.5 u/kg in cases of severe insulin resistance) depending on patient size and degree of glycaemia'),nl,

write('Adults of normal weight may be started with 20-25 u/d of intermediate acting insulin and increased to maintain a blood sugar level of 80-120 mg/dl'),nl,

Declaration

I declare that the thesis is my original work and has not been presented for a degree in any other university.

_____

Tesfaye Fufa

October, 2015

This thesis has been submitted for examination with my approval as university advisor.

_____

Dereje Teferi (PhD)