ADDIS ABABA UNIVERSITY
SCHOOL OF PUBLIC HEALTH
AND
SCHOOL OF INFORMATION SCIENCE

M.SC IN HEALTH INFORMATICS

DEVELOPING A PREDICTIVE MODEL FOR PRE- DIABETES SCREENING BY USING DATA MINING TECHNOLOGY

BY

BEZAHEGN ZERIHUN

June, 2017

ADDIS ABABA UNIVERSITY
SCHOOL OF PUBLIC HEALTH
AND
SCHOOL OF INFORMATION SCIENCE

M.SC IN HEALTH INFORMATICS

DEVELOPING A PREDICTIVE MODEL FOR PRE- DIABETES SCREENING BY USING DATA MINING TECHNOLOGY

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Health Informatics

BY

BEZEHEGN ZERIHUN

June, 2017

ADDIS ABABA UNIVERSITY
SCHOOL OF PUBLIC HEALTH
AND
SCHOOL OF INFORMATION SCIENCE

M.SC IN HEALTH INFORMATICS

DEVELOPING A PREDICTIVE MODEL FOR PRE- DIABETES SCREENING BY USING DATA MINING TECHNOLOGY

BY

BEZEHAGN ZERIHUN

Name and Signature of Member of Advisors

3. Dr. Million Meshesh (PhD) Signature_____ Date _____

2. Dr. Assefa Seme (PhD) Signatures_____ Date _____

Name and Signature of Member of Examiners

1 Dr. Marta Yifru (PhD) Signature _____ Date_____

2 Dr. Eshetu Girma (PhD) Signature _____ Date_____

DEDICATION

I would like to dedicate this thesis to my parents, brothers and sisters for all their love and support

ACKNOWLEDGMENTS

I would like to thank all those who have helped me to accomplish this project. First, I gratefully express my deepest thanks to the almighty God for helping me the accomplished the paper and who added years to my life, Glory to God. Next, I would like to extend my deepest thanks to my advisors Dr. Million Meshesha (PhD) and Dr. Assefa Seme (MD, PhD) for their wonderful and unreserved assistance in every step of this study.

I would like also to extend my thanks and appreciation to Adare hospital management and diabetes unit staff specially Mr. Fikru Tesfaye (CEO Adare hospital) Dr. Eyob, Sr. Adanech and Mr. Genene, for their cooperation and willingness to provide me with all valuable information .

I would also like to thank Addis Ababa University, School of Information Science and School of Public Health and Hawassa city administration health department for financial support and overall facilitation of the research from the beginning until the end.

Finally I would like to thank importantly my all family members specially (Afire, mum) brothers and sisters, friends. Special thanks for Meseret Ayano. (SIS), Markos Buta. Mr. Erimias Tenaw, and Desalegn Dabaro for their continuous and credible effort to realize my work. My classmate's special Sr. Bethlehem lemma and Atiklt Michael for their comments, constructive ideas, moral support and understanding during the time I solely devoted to the study.

Table of Contents

List of Figures

Lists of Tables

List of Acronyms and Abbreviation

BG	Blood Glucose
BMI	Body Mass Index
CSA	Central Statistical Agency
CVD	Cardiovascular Disease
DM	Data Mining
FMOH	Federal Ministry of Health
FBS	Fasting Blood Sugar
HSTP	Health Sector Transformation Plan
HSDP	Health Sector Development Program
HMIS	Health Management Information System
IDDM	Insulin Dependence Diabetes Mellitus
IDF	International Diabetes Federation
KDD	Knowledge Discover in Data Base
KDP	Knowledge Discovery Process
LADA	Latent Autoimmune Diabetes In Adults
MIS	Management Information System
NCD	Non Communicable Disease
NIDDM	Non- Insulin Dependent Diabetes Mellitus
NN	Neural Network
SNNPR	Southern Nation Nationality Regional State
SVM	Support Vector Machine
T2DM	Type Two Diabetes Mellitus
TB	Tuberculosis
WEKA	Waikato Environment for Knowledge Analysis

Abstract

Introduction *- Diabetes is one of the most common non-communicable diseases. That has a significant contribution of increased morbidity, mortality and admission rate of patients in both developed and developing country. The burden is also very enormous in Ethiopia with estimated 1.4 million in World Health Organization country profile report (2014); even this doesn't included pre-diabetes and undiagnosed cases. International Diabetes Federation report an estimated 83.8% of all cases of undiagnosed diabetes mellitus are in low- and middle-income countries. Therefore early screening, diagnosis and prompt treatment are needed to prevent comorbidity and mortality, delay the onset of disease, and reduce serious complication and permanent damage.*

Objective: *The aim of this study was to develop a predictive model for screening of pre-diabetes patient using data mining technology.*

Method: *This study conducted in Adare general hospital in Hawassa city, south Ethiopia. The methods used for mining, Cross-Industry Standard Process of Data Mining which contains six phases such as problem understanding, data understand, data preparation, model building, evaluation and deployment was used. In general, 4529 of age > 20 years visiting diabetic unit for general medical examination and follow up were included from January to March 2017. Designed template was used for data collection. For data pre- processing was used Microsoft Excel and WEKA open source software for mining.*

Results and discussion: - *The study has revealed that the model constructed PART with all attributes registers the highest accuracy of 96.78% as compared to J48 decision tree which was 93.66%. The finding of the study clearly presents that screening of diabetes and pre diabetes patient. Based on result of prediction designed project prototype model that predict whether the positive risk of diabetes or not based on this result patients should link further investigation or provide council for future the way to prevent or delay on set of diabetes.*

Conclusion: - *Generally, the prototype system serves as a guideline, diabetic screening to support early detection of patient. The initial feedback from health works has been extremely positive. Hence the developed prototype system achieves a good performance and meets the objectives of the project.*

Recommendation: *- based on finding of project forwarded recommendation for respective stockholders.*

CHAPTER ONE

INTRODUCTION

1.1 Background

Diabetes is one of the most common non-communicable disease (NCDs) that has significantly contributed to increased morbidity, mortality and admission rate of patients in both developed and developing country(1). It is assuredly one of the most challenging health problems in the 21st century that evidently is epidemic in a large number of populations in developing countries. Over the past decade, diabetes prevalence has risen faster in low- and middle-income countries than in high-income countries. Nature of diabetes disease is an asymptomatic in early stage and can remain undiagnosed for 9 to 12 years. American diabetes Association report (2014) one in six affected at age of 60. In recently prevalence indicated epidemiological transition of disease to younger age groups(1, 2).

The transition imposes more constraints to deal with the double burden of infective and non-infective diseases in a poor environment management and nutrition intervention. Literally diabetes is disease of developed nation. This is misunderstanding thought because rapid growth of urbanization, poor culture of the physical activity and high prevalence of the obesity are contributors of the increase diabetes (3, 4).

The disease has variable group of clinical symptoms and sequences. It describes a metabolic disorder of multiple etiologies' characterized by chronic hyperglycemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both.(5).

Diabetes affects all segment of population irrespective of age and sex(6). Diabetes of all types can lead to complications that can increase the overall risk of dying prematurely. Possible complications include heart attack, stroke, kidney failure, leg amputation, vision loss and nerve damage. In pregnancy, poorly controlled diabetes increases the risk of fetal death and other complications (7).

Management of complication it requires high cost, increase time of the hospitalization and poor treatment outcome. Diabetes related complication have a 2-fold increased risk of stroke and

cause lower extremity amputations in diabetes related 10 times more common in people with diabetes than in non-diabetic individuals in developed countries and more than half of all non - traumatic lower limb amputations are due to diabetes . On the other hand they would affect people and lead to various complications like blindness, gross metabolic disorder and sexual dysfunction(8).

End stage diagnosis Diabetes mellitus may present with characteristic symptoms such as thirst, polyuria, blurring of vision, and weight loss; in most severe forms, ketoacidosis or a non-kenotic hyperosmolar state may develop and lead to stupor, coma and, in absence of effective treatment, to death (9).

People with diabetes require at least 2-3 times the health care resources, which do not have diabetes, and it accounts for up to 15% of national healthcare budgets in the developed countries. If diagnosis is not done early the patient undergoes different acute and chronic complication. This inclines to raise the management cost of patient. IDF in 2014 report announced diabetes caused 4.9 million deaths and costs USD 612 billion in healthcare spending annually(10).

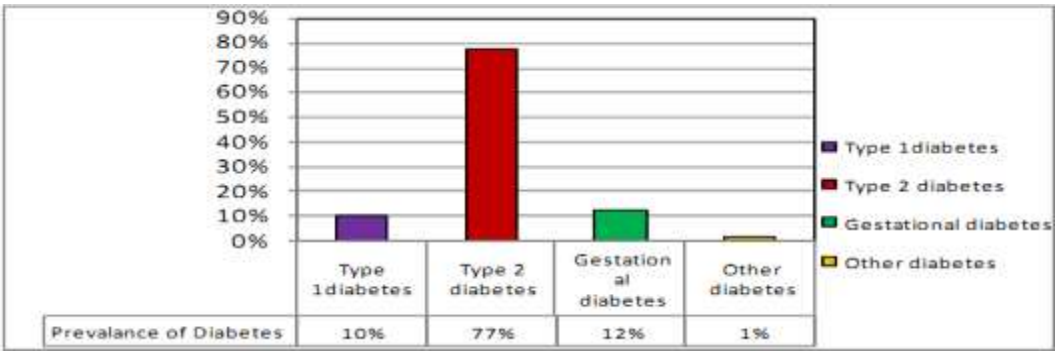

Figure 1: Prevalence of different types of diabetes source http://www.somersetintelligence.org.uk/diabetes.html

The above figure describes the prevalence and type of diabetes the figure indicates that type2 diabetes is high prevalence and contribution of the two gestational diabetes and type1 account 22% others 1% proportion.

1.1.1. Diabetes Mellitus in Ethiopia

Developing countries including Ethiopia experiences a heavy burden of infectious disease mainly attributed to communicable infectious diseases and nutritional deficiencies. Currently non-communicable diseases are also major health problem in Ethiopia. According to the World Health Organization-NCD country profile of 2014, Non-Communicable Diseases are estimated to account for 30% of total deaths in Ethiopia. Diabetes are third leading cause hospital admission and 4th leading cause of death (11). This characterize high prevelence of the nutritional deficiencies and shortage of both access and quality of the health care services coverage, low maternal and child health and nutritional accounts 60% and accident injury accounts 10% of total cause the reaming 30% is non- communicable disease. NCD third leading cause of morbidity and 4th leading cause of mortality in Ethiopia(12).

1.1.2 Data Mining in Health Care

Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information from large data(13). It provides a user- oriented approach to novel and previously identify patterns. The discovered knowledge can be used by the healthcare administrators to improve the quality of service and can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Data mining technology assist health care researchers to extract hidden knowledge which leads to improve health care delivery, support technical and managerial decision-making by introducing clinical decision support system and artificial intelligent. Its bench mark of prevention strategy, enhance chronic disease management and improve quality of care of patient(14).

The process of the decomposing of medical diagnoses and creates additional choice to introduce periodic screening and enhance early detection of patient based on risk assessment of clinical and epidemiological data helping of international standards of IDF and WHO agreed criteria(15).

Definition of Terms

Diabetes:-Diabetes mellitus (DM) is a group of diseases characterized by high levels of blood glucose resulting from defects in insulin production, insulin action, or both**.**

The term diabetes mellitus describes a metabolic disorder multiple aetiology characterized by chronic hyperglycaemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both.

Screening

"Screening is the process of identifying those individuals who are at sufficiently high risk of a specific disorder to warrant further investigation or direct action."

"It [screening] is systematically offered to a population of people who have not sought medical attention on account of symptoms of the disease for which screening is being offered and is normally initiated by medical authorities and not by a patient's request for help on account of a specific complaint. The purpose of screening is to benefit the individuals being screened." *WHO & IDF 2003*

1.2. Statement of the Problem

According to WHO and IDF report currently globally, an estimated 422 million adults were living with diabetes in 2014. The global prevalence of diabetes has nearly doubled since 1980, rising from 4.7% to 8.5% in the adult population. Diabetes is fourth leading causes of NCD deaths in 2012 were: cardiovascular diseases leads to 7.5 million deaths, cancers 8.2 million, respiratory diseases, including asthma and chronic obstructive pulmonary disease 4.0 million, and diabetes 1.5 million deaths. These four major NCDs were responsible for 82% of total non-communicable deaths(16).

WOH and IDF survey global report (2015) Notify that Diabetes Mellitus prevalence has been increase alarming rate. WHO fact shit (2015) among global prevalence 77% of people with diabetes live in low- and middle-income countries, and the socially disadvantaged country are the most vulnerable to the diabetes disease related complication. The developing country human

and financial costs of management of diabetes are high and escalating from time to time. Developing countries especially upper level country Nigeria and china other line up with this global trend. Other research report the trends of diabetes disease gradual move towards developing countries, because principle of western diet in many societies in developing countries the same due to globalization, expansion of urbanization , poor physical activity or sedentary way of life and poverty(10).

WHO report (2014) inform that currently 12.1 million people were estimated to be living with diabetes in Africa, and this is projected to increase to 23.9 million by 2030. It is estimated result based on WHO country profile, currently 1.4 million people are living with the diabetes in Ethiopia this projects 2.7 million in 2030 the prevalence of diabetes is greater than prevalence of HIV/AIDS less than 1milion in Ethiopia 2014. Prevalence of diabetes figure is not show pre diabetes and undiagnosed diabetes according to IDF report An estimated 83.8% of all cases of undiagnosed diabetes mellitus are in low- and middle-income countries(16, 17).

Diabetes is one of the leading causes of chronic complication such as visual impairment (blindness) in developed countries. People with diabetes require at least two to three times the health-care resources compared to people with no diabetes, and diabetes care may account 15% of national health care budgets in developed country. In addition, the risk of tuberculosis is three times higher among people with diabetes(7).

Different literature suggested that detecting of diabetes 80% of disease progress is established on Clinical suspicion is confirmed by doing a laboratory assessment of the oral glucose or sugar level in the patient's blood sample. These methods is not feasible for screening, because cost needs skilled man power and time consuming so not accessible all segment of population. According to survey of CSA 1997 data 84% people living in rural area, health institution highly concentrated on central part of city. On the other hand, current health care setup busy outpatient setting. The shortage of high trained health care providers is acute problem in Ethiopia. Today ultimately raise cost of patient health care service for treating non-communicable disease like diabetic patients and improve the quality of care(18) (19).

Data mining is important technology to build a model that predicts the screening status of pre-diabetes patient thus enabling early detection prevent complication in Ethiopia. This will farther

help to protect the health of individual from morbidity, mortality, reduce diabetes related complication and the reduce treatment cost due to early detection; reduce economic burden individual and country already scarce resource of the healthcare system in Ethiopia.

1.3 Objective

1.3.1 General Objective

The general objective of this study is to develop pre- diabetes screening model using data mining technology, which can help the effort of early detection of diabetes

1.3.2 Specific Objectives

The following specific objectives are attempted in this study

- To collect and code diabetic and pre-diabetic patient data
- Identify and decide relevant attributes for mining
- To identify and compare best algorithms models for screening diabetes.
- To develop a prototype for diabetic screening
- To evaluate the user acceptance of the prototype.

1.4 Scope and Limitation of the Project

The scope of this project is to apply data mining technique to construct a prototype diabetes screening model. The project was geographically limited to SNNPR in Hawassa area; the reason behind is because of financial and time constraints. But developed model works all hospital and Health centers. The data covers the period from January to March 2017; which amounts to 4529 records. Thus, the research project can be considered as primarily pertinent to the Hawassa city health department and Adare hospital to make decision regarding implementing the project.

This study mainly focus on type one and type two which account 89 % of all cause of diabetes because all others are rare cause and require different research design such a causes of diabetes account 5% total cases like gestational diabetes, drug induced diabetes is not included because of research design, time and financial constraint.

1.5 Significance of the Project

The main aim of this research project is to develop accurate, user friendly, simple to use model to support medical practitioners predicting the pre diabetes screening of not yet diabetes or undiagnosed diabetes. The early prediction of disease gives a warning about the level of risk and that arise due to diabetes, where treatments and preventative action can initiate and the patient to extend the period of patient's enjoying healthy life.

The project output enables to reduce human or subjective definition for the time of screening. To minimize health care professional individual definition for diabetic screening endorses the standard case definition through developed screening model.

For patients

- Prevent diabetic related complication by early detection
- Reduce cost of treatment
- extend healthier life and increase production

For medical practitioners

- Reduce subjective and technical definition during screening.
- Reduce burden of diagnosis and examination by screening non- illegible population.
- Easy to screening procedure by helping computer

For health facility

- Reduce patient our load in outpatient setup
- Improve customer satisfaction
- Save resources both human financial resource

In The Health System

- To introduce new approach for diabetes screening
- To reduce cost of screening with integration with existing services

- Important for health care planers and policy makers to pass evidence based decision and foundation for the diabetes researches.

1.6 Ethical Consideration

For this research project I has been use patient data for training and test set of the algorism, the research project primarily for academic achievement on MSc in Health informatics. Ethical clearance was obtained from Research and Ethics committee of the School of Public Health of Addis Ababa University. And permission letter also be produced from respective regional health Bureau and City health department. During time of data collection the investigator maintaining the confidentiality of participant. During data collection the information sheet contains the objective of the project, maintain confidentiality, the template is attached to annex. All other ethical issues are addressed as important component of the research

1.7 Dissemination of the Result

The result of the research will hopefully be presented at board of examiner students conference in Addis Ababa University, Hawassa city administration and Adare general hospital, national conference of Ethiopian public health association, SNNPR health bureau.

1.8 Organization of the Paper

This research project report is organized in to six chapters. Chapter one is introduction and it cover statement of the problem, objective, significance of project and methods of the research Chapter two reviews the pertinent literature and other related governmental published and unpolished document and related works. Chapter three is methods of the research, design of data, data collection coding procedure and detailed of stapes of data mining. Chapter four presented business understanding, data understanding, and data preparation for experimentation. Chapter five experiments on each algorism compare each algorism and select the best algorism based on selected rules develop prototype and test user acceptance. Finally chapter six forwards conclusions and recommendation of the research project.

CHAPTER TWO

Literatures Review

2.1 Background of Non-Communicable Disease

Non communicable diseases are sweeping the entire globe; with an increasing trend in developing countries the high burden of communicable disease is because of poor sanitation and nutritional related disorder where, the transition imposes more constraints to deal with the double burden of infective and non-infective diseases. The reason of increasing of people with diabetes suggested different factor that expansion of urbanization and sedentary life of population, aging increasing and prevalence of obesity and physical inactivity. Great contribution for prevalence of diabetes(20).

2.2 Background of Diabetes Mellitus

The diabetes mellitus disease is a metabolic disorder of multiple causes characterized by chronic hyperglycemia with disturbances of carbohydrate, fat, and protein metabolism resulting from defects in insulin secretion, insulin action or both. A chronic metabolic disorder of multiple etiologies is assuming epidemic Proportions both developed and developing countries (21). WHO report of 2011 declares that, diabetes has become major health problem in developing countries and has been found in a wide variety of atypical forms. Its burden is huge in developing countries due to lack of basic means for reaching diagnosis and a reasonable glycemic control. In most developing countries there is no diabetes mellitus epidemic control strategies and because the nature of the disease asymptomatic unless destruction of the insulin secretion cells or resistances of utilization glucose they do not any sign and symptoms. So finally the patient comes with one or two diabetic related complication. The developing country majority of people are poor health seeking behaviors in low resource countries because of inaccessible health care facility, lack of skilled man power and medical supply results in low quality health care.(22)

2.3 Diabetic Related Complication

Patients with Diabetes Mellitus are at increased risk of developing chronic complication. Most of the complications are permanently damage organ or system of the body. Such complication includes the following, heart disease patient may die heart disease secondary to diabetic, kidney

failure or disease, neuropathy or defects nerve function due to this non- emergency amputation. People with diabetes carry a risk of amputation that may be more than 25 times greater than that of people without diabetes. However, with comprehensive management, a large proportion of amputations related to diabetes can be prevented Stroke and loss of vision. Diabetes is the leading cause of blindness and visual impairment in adults in developed countries(21).

2.4 Prevention Strategies

According to WHO technical report(22). Suggested to implement the following effective prevention strategies existed at three deference levels as

- Primary prevention – covers all activities aimed at preventing diabetes from occurring insusceptible individual or population through modification of environmental and behavioral risk factor/ determinants or specific interventions for susceptible individuals.

There are two ways of primary prevention strategies

I. Activities targeted at reducing the frequency or level of causal risk factor for development of diabetes in whole population or a group of individual particularly that at high risk in general community based prevention strategies.
II. Activities targeted at prevention specific individual who are already manifesting early markers of disease process from developing the full clinical expression of disease of diabetes this could include intervention strategies (Pharmacological and non-pharmacological in individual with abnormal glucose tolerance or other metabolic abnormality, immunological or other markers of beta cell distraction.

- Secondary Prevention- Covers all activities such as screening which aims at early detection of diabetes and prompt the effective management of the condition with propose of reversing the condition or extending the propagation of disease. This include many strategies integrated at the detection as yet undiagnosed cases of diabetes, aging activities can be targeted acted on the target group or population at risk.
- Tertiary prevention – is any measure undertaken to prevent complication and disability due to diabetes i.e. to prevent or delay the negative health consequence of diabetes or diabetes related complication among individuals ,who have already develop the disease in particularly this means early detection, effective management, education and metabolic control as well as

the correction or redaction of major risk factors for specific disorder among people with diabetes(23).

2.5 Overview of Data Mining

Health care industry is the fast-growing, with large amount of data, collected and stored in numerous databases. Manual and digital data collection and storage tools have far exceeded our human potential for comprehension without powerful tools. Data mining is rapidly growing and successful in a wide range of applications such as analysis of organic compounds, financial forecasting, healthcare and weather forecasting. Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Applications in healthcare include analysis of quality of health care delivery, tolls for policy-makers towards sound decision making and prevention of committed treatment errors, early detection, prevention of diseases and preventable hospital deaths, more value for money and cost savings, and detection of fraudulent insurance claims (24).

Another factor motivating the use of data mining applications in healthcare is the realization information that can generate very useful to all parties involved in the healthcare industry. For example, Healthcare insurers detect fraud and abuse, and healthcare providers can gain assistance in making decisions process, for example, in customer relationship management. Data mining applications also can benefit healthcare providers, such as hospitals, clinics and physicians, and patients, for example, by identifying effective treatments and develop and utilized best practices(25).

On the other hand, as the writers adopt the convention that data mining refers to the act of extracting patterns from data (be it automated or human-assisted). However, many steps precede the data mining methodology: retrieving the data from a large warehouse (or some other source); selecting the appropriate subset to work with deciding on the appropriate sampling strategy; cleaning the data and dealing with missing fields; and applying the appropriate transformation, dimensionality reduction, and projections. The data mining step then fits models to extracts patterns from the preprocessed data. However, to decide whether this extracted information does represent knowledge or one need to evaluate this information visualize it, and finally consolidate it with existing (and possibly contradictory)

knowledge. Obviously, these steps are all on the critical path from data to knowledge. Furthermore, any one-step can result in change in the preceding or succeeding steps, often requiring starting from scratch with new choices and settings(26).

The data mining process model helps organizations to better understand the KDP and provides a roadmap to follow while planning and executing the data mining project. This in turn results in cost and time savings, better understanding and acceptance of the project results. Such processes are nontrivial and involve multiple steps, reviews of partial results, possibly several iterations, and interactions with the data owners. The knowledge discovery process models usually emphasize independence from specific applications and tools; they can be broadly implement in different such as industrial and academic model(27).

2.5.1 Knowledge Discovery in a Database

The data-mining field currently relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data in the data-mining step of the KDD process(28).

The KDD process of data preparation, data selection, data cleaning, incorporation of prior knowledge, and proper interpretation of the results of mining ensure that useful knowledge is derived from the data. An important notion of "interestingness" is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, simplicity and understandability. As a matter of fact, knowledge in this definition is purely user oriented and domain specific and it is determined by whatever function and threshold the user chooses. The roles of interestingness are to threshold the huge number of discovered patterns and report only those which may be of some use(29).

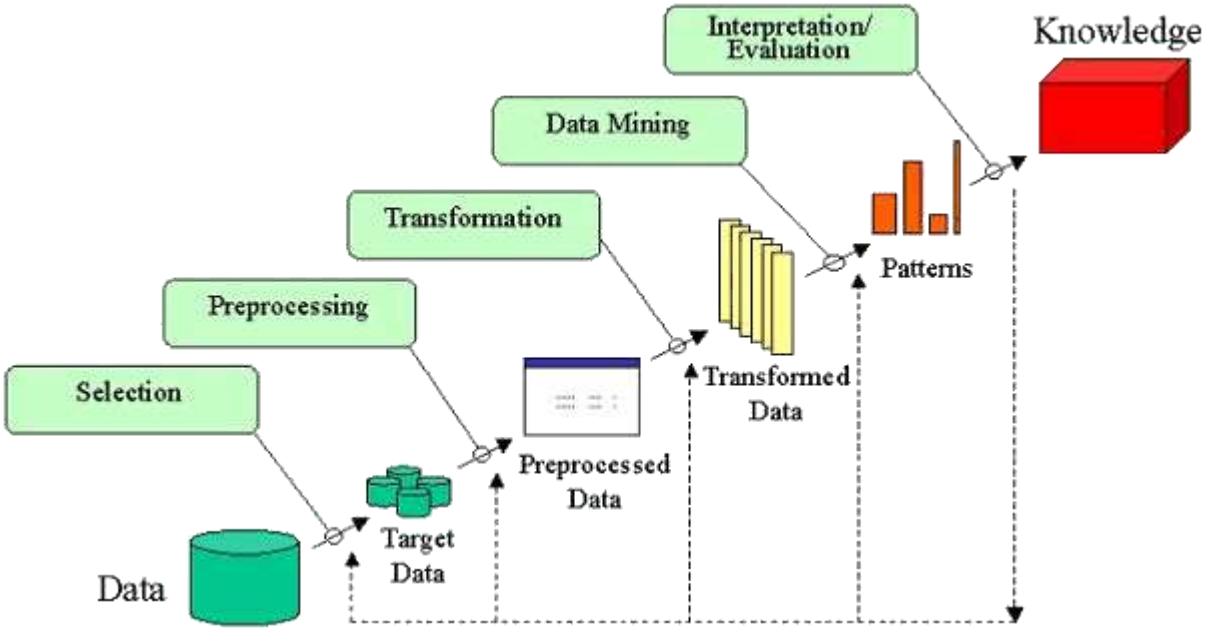

Fig 2.1 The Knowledge discover in database (KDD) source (MIT Press, Menlo Park, CA, 1996)

In view of machine learning process or KDD refers to the overall process of discovering useful knowledge from data. The distinction between the KDD process and the data-mining step (within the process) is a central point of this article. The additional steps in the KDD process, such as data selection, preparation and cleaning and incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, is essential to ensure that useful knowledge is derived from the data (28).

Knowledge discovery concerns the entire knowledge extraction process including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain area. 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 data mining component of KDD currently relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data in the data-mining step of the KDD process(30).

Lastly, when as writers explains encounter patterns within a database the researchers state the findings (patterns or rules) as data mining, information retrieval or knowledge extraction and so on. The term data mining is used mostly by statisticians, data analysts and the management information systems (MIS) professionals. The difference between DM and KD is that the latter is the application of different intelligent algorithms to extract patterns from the data whereas KD is the overall process that is involved in discovering knowledge from data. There are other steps such as data preprocessing, data selection, data cleaning, and data visualization, which are also a part of the KDD process. Many people treat DM as a synonym for another popularly used term, KD from Data, or KDD. Alternatively, others view DM as simply an essential step in the process of KD. Hence, in the definition as the writer adopt, DM is just a step in the overall KDD process(29).

2.5.2 Cross Industry Standard Process for Data Mining (CRISP-DM)

Cross-Industry Standard Process for Data Mining or CRISP-DM majority of sharing in data mining. This Proposes of CRISP-DM methodology for mining: a multi-step iterative process. It consists of six phases such as business understanding special the way of the business perspective is critical because it identifies the business objectives and, thus, the success criteria of data mining projects. (25)

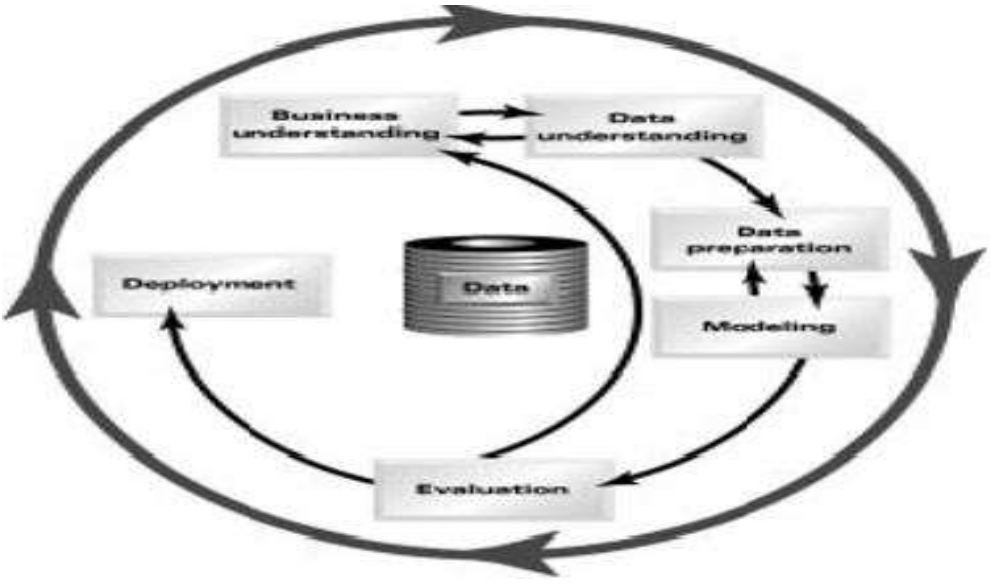

Fig 2.2 Phases of the CRISP-DM reference model Source: (Chapman et al, 2000)

1. **Business understanding**: - This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM Problem definition and a preliminary plan designed to achieve the objectives.
2. **Data understanding:** This phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
3. **Data preparation:** The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.
4. **Modeling**: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type.
5. **Evaluation:** Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives.
6. **Deployment:** Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it(24).

2.5.3 Hybrid Models

The development of academic and industrial models has led to the produce of hybrid models that combine aspects of both. One such model is a six-step KDP model developed by Cios et al. It was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include providing more general, research-oriented description of the steps and introducing a data mining step instead of the modeling step, Further description of the six steps see figure 2.3 of the model follows(31),

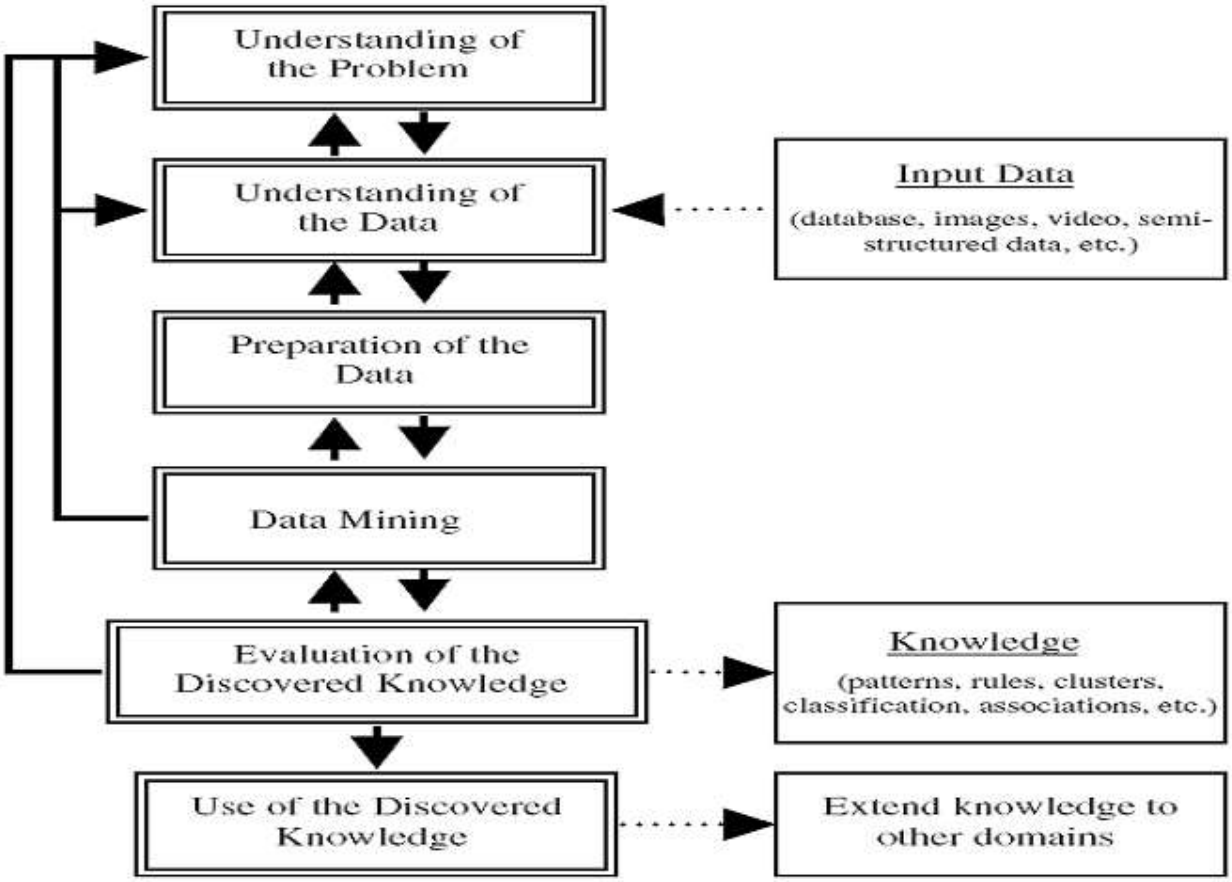

Fig 2.3 Hybrid Process Model Source ( )

1. **Problem domain understanding**- This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

2. **Data understanding-** This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

3. **Data preparation**- This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data generalization). The end results are data that meet the specific input requirements for the DM tools selected in Step 1.
4. **Data mining-** Here the data miner uses various DM method such as classification, clustering and association rule discovery to derive knowledge from preprocessed data as per the objective of the study.
5. **Evaluation of the discovered knowledge-** Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.
6. **Use of the discovered knowledge**- This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

2.6. Data Mining and Health Care

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve health care and reduce costs. Some experts believe the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending.

This could be a win/win overall. But due to the complexity of healthcare and a slower rate of technology adoption, data mining techniques has proved for early prediction of disease with higher accuracy in order to save life and reduce the treatment cost (32). IT solution accelerates

implementation of health care crevice decreases the running cost of patient management. Incorporating different information technologies (ITs) into the healthcare system of developing countries is not all about modernizing the health system but it is about saving life by facilitating communication, practicing evidence based decision, using previews stored data, experience sharing from success history, incorporating e-learning to remote health professionals, use it as a medium to access recent healthcare information, data handling and processing activities among staffs and handling of patient care(33).

Treatment effectiveness: Data mining models can aid in evaluating the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective. For instance, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective. Successful standardized treatments for specific diseases can also be identified using data mining. Other data mining applications related to treatments include associating the various side-effects of treatment, identifying common symptoms to aid diagnosis, determining the most effective drug compounds for treating patients who respond differently from other patients to certain drugs, and determining proactive steps that can reduce the risk of drug side-effects(34).

In recent years, data mining has been used widely in the areas of science and engineering, bioinformatics, genetics and medicine. It is a collection of algorithmic ways to extract informative patterns from raw data and considered as an example in health care setting. It plays an important role in tackling the data overload in medical informatics and provides a user oriented approach to novel and hidden patterns in the data and its applications can be developed to evaluate the effectiveness of medical treatments and to identify the behaviors of disease by using extracted pattern(35).

Data mining methods in the medical domain are helping due to the increasing effectiveness of classifications that help the doctors especially in decision making. Data mining algorithms had been applied to classify diabetic patients. It is used for non-laboratory attributed to classify the diabetes by applying suitable algorism. Effective screening methods have been proposed taken to predict the disease at the earliest and control deterioration of cases. using various tools, PART rules induction and Modified J48 decision tree(36).

2.7 Related Works

In a study conducted Y, Hongmei is three methods of data mining. The study described the construction of a clinical decision support system to predict the presence of myocardial infraction in a cohort of 4,770 patients presenting with acute chest pain at two university hospitals and four community hospitals.

15 attribute selected based on the patient's symptoms and signs, the clinical decision support system had similar sensitivity (88.0% versus 87.8%) but a significantly higher specificity (74% versus 71%) in predicting the absence of myocardial infarction when compared to physicians' decisions if the patients were required to be admitted to the coronary care unit. If the decision to admit was based solely on the decision support system, the admission of patients without infarction to the coronary care unit would have been reduced by 11.5% without adversely affecting patient outcomes or quality of care(37).

Another research Beck Huain and Y, huajo 2008, is focused on diabetes related complication prevention. About 30 to 80 percent of type 2 diabetic cases remain undiagnosed suggested that data mining, using decision trees, type2, with different degrees of prevalence in society. It was recognized by an asymptomatic phase between the real onset of diabetic hyperglycemia and clinical diagnosis within 4-7 years

Methods of data collection from visitors consider as risk factor data collection period from 2009-2011. The attributes selected personal and epidemiological linkage when patient status as the time of visited health facility. Element of selected attribute are obesity or overweight, history of diabetes in first-degree relatives, hypertension in pregnancy, privies history of gestational diabetes, history of abortion, stillbirth, and birth of a child more than 4 kg, and background of patient and epidemiological data Features including age, sex, family history of diabetes, and body mass index (BMI)

The investigator used technique of decision tree and J48 algorithm for developing the decision tree in WEKA (3.6.10 version), the accuracy of model checked precision and accuracy of the model was 71.7 and 97.6 percent, respectively. Researcher concluded that the developed model using the decision tree for the screening of T2DM that does not require laboratory tests for

diagnosis. The researcher suggest used J48 algorithm and decision tree model proposed is for three reasons: easily access the risk of patient, features applied to primary screening, excluding those the risk such as plasma glucose for the main diagnosis of T2DM and capability of the decision trees for T2DM screening. Although the exclusion of diabetes laboratory diagnostic tests features lowered the sensitivity and precision of the model proposed compared with models suggested in the literatures(38).

Another study is the objective of the predicting chronic asthma: Razak & Bakar (2006) have conducted a study focusing on mining association rules from asthma patients profile dataset. The purpose of the study is to identify attributes that affect asthma patients.

The asthma patients profile dataset in this study consists of 16,384 records and 118 variables in various formats. These attributes are grouped into demographic attributes and asthma related attributes. The mining method used involves data preparation phase and association rules mining phase. Understanding the nature of the dataset, identifying data types and formats, identifying incomplete data, analyzing data distribution, and discretizing data are the important stages needed in order to systematically preprocess the data. As a result of data preprocessing and cleansing only 31 attributes are left to generate association rules. The association rules mining phase uses A priori Algorithm. Determining training and testing datasets, determining threshold values, mining association rules and association rules analysis are involved during the implementation of the rule mining(39).

Selam,A conducted a project on the measles outbreak across different region in Ethiopia. The methodology building predictive model using data mining technique for this research was a hybrid six-step Cios KDP. It had six basic steps. Model build by 13 selected attributes for building predictive model. Investigator experiments have been carried out with two classification algorithms, the decision tree and naïve Bayes Models; they disagree on the classification of several outbreaks. The classifier has 86.8% sensitivity which shows that the model has acceptable capability of recognizing the true positive value of the class "yes" and 99.7% specificity. The second experiment used 9 attributes and scored the best accuracy of 93.31% with 70% split test option from the other experiments. Experiment three scored the most accuracy with both test option. Selected algorism recommends region based Measles outbreak prediction(40).

Shegaw conducted study focus on the investigate the potential applicability of data mining technology to predict the risk of child mortality based up on community based epidemiological datasets. The researcher used neural network and decision tree methods. In order to building test the models. Using the neural network approach, the best model was identified for the training made by using the default parameters build from 9 attribute. This model had an accuracy rate of 93%.

This classifier resulted with an accuracy of 95% on training cases and it achieved 95% accuracy on test cases. The researcher concluded that this research work have proved the potential applicability of data mining technology to predict child mortality patterns based on demographic, parental, environmental, and epidemiological factors. The encouraging results obtained from both neural networks and decision trees indicate that data mining is really a technology that should be considered to support child health care prevention and control activities at the district (41).

In my current knowledge, no previous researches have been done to predict the early screening of diabetes disease by applying data mining techniques in Ethiopia. But, some scholar had been done data mining technology on the other disease, such as incidence pattern prediction, occurrence of measles epidemic distribution on region and under five children mortality in Ethiopia.

CHAPTER THREE

RESEARCH DESIGN AND METHODOLOGY

3.1 Research Project Design

There are abundant data mining techniques presented today with their appropriateness to be applied in different health care areas; Such as clinical decision support system, improve quality of care of patient and health care research and development. The data mining techniques used to predict the occurrence of different health care problems like epidemic and community based and the facility based screening program both developed and developing country.

Developing country like Ethiopia is shortage of resources. The health programs assisted by technology results save resources increase efficiency of the program. While introducing this new approach for diabetes screening program can be based diabetes and pre diabetic patient data were used decision tree and rule induction generate classifier which is easy and simple to implement evaluate the trends of the program.

The CRISP-DM technique is followed to explore the application of data mining diabetes screening all eligible groups. This model was chosen since it exhibits all the advantages of well-known and used methodology called CRISP-DM and provides a more general, research-oriented description. Data mining technology provides a user- oriented approach to novel and hidden patterns in the data. There are six stapes (33). .

3.1.1Business Understanding

A model was needed to identify and predict undiagnosed diabetes the contribution of the designed attribute for prevalence of diabetes. Support preventive measures before it causes diabetes related complication harm to the society. The researcher used literature from national and international report for reference and consultation with domain experts. To achieve the goal of this research an attempt is made to adopt diabetes screening template from WHO guidelines and Iran ministry of health screening tools.

3.1.2 Data Understanding

The primary source of data for this research is diabetic and pre-diabetes patient data from patient folder. This method no goes far recorded data missing value and full of inconsistences. The researcher change way data collection data. Data was collected from diabetic clinic and other general medical checkup chronic follow-up units from January to March 28/2017 for used for building the model. The collected data was in patient folder, it contains a total of 4529 records about patients from Hawassa city and surrounding rural district. The dataset contains both numeric and nominal values. The data contains information about the location, genetic history of subject, basic information of visitor like age, sex, and physical and behavioral risk factor, obesity, history of physical activities and comorbidity with the other chronic disease.

3.1.2.1 Data source

The source of the data used to undertake this research was patients' data taken from Adare general hospital. Adare general hospital is a public hospital which is found in Hawassa city administration. The main reason to select this hospital is, there are high number of diabetes and chronic patient flow, well originated diabetes units and other facility concerning the problem when compared with other hospitals and through discussion with the domain experts from regional health Bureau and Hawassa city health department.

3.1.2.2 Data Collection Methods

Data collection methods using this project primary data by direct patient interview, the investigator chose these methods the problem of data incompleteness due to missing value. Time of data collection was from January to March 2017, data collection methods structured designed template see annex,

3.1.2.3 Sampling Methods

The sampling methods for this research project are Convenience sampling techniques the reason chose this sampling technique, Convenience sampling is used in exploratory research where the researcher is interested in getting an inexpensive approximation of the truth. This nonprobability method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a random sample. The study subject is

defined population diabetes and pre diabetes patient the yield of simple size decided by the required time of data collection. Total sample size 4529 for three months from January to march2017 subjects to build model for diabetes screening model.

3.1.3 Data Preparation

This is one of the crucial steps to produce quality dataset used for modeling by WEKA software. The following steps were undertaken in the preprocessing stage; data coding, data cleaning, attribute selection, and data transformation. Data interring, trimming unnecessary value and coding or modeled some derived attribute. In order to correct the errors identified through observation from the preprocessing stage. Finally, the dataset to ready for the data mining process contains 17 attributes (including class) on the total of 4529 instances.

3.1.4 Modeling

To build a predictive model from the cleaned data, WEKA 3.6.9 version open source data mining software was used. WEKA is a tool containing numerous machine learning algorithms that can be applied to achieve the objective of this research project. It supports standard of data mining tasks; more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection and also this software is platform independent classifier. The researcher chooses WEKA software constricting a predictive modeling using decision tree and PART rule induction algorithm. These algorithms were proved to be important when applying them in healthcare data recommended because of easy to interpret result. Therefore J48 algorithm and PART algorithms were applied on the diabetes screening data to come up with the predictive model for the risk of diabetes screening. Based on the result design model screen patient on the top of these start intervention strategies to break the epidemics or delay developing diabetes.

3.1.5 Evaluation

Before preceding the final model development, evaluate Performance and accuracy of the model created by the J48 Decision tree and PART rule induction. The methods relevance checked using confusion matrix, ROC curve, 10 folds cross validation and prepared dataset spited with 70% split for training and 30% for testing.

3.1.6 Deployment

The model construction is not the end of project. The purpose to produced knowledge from the data by classification methods; after the experimentation was selected algorisms to extracted rule to build best predicative diabetes screening model.

3.2 Architecture

The major components of any data mining system architecture are the guideline of the data modeling for prediction data source, data mining engine, pattern evaluation module, and graphical user interface. The data mining process is identifying the most effective model of diabetes screening in adult people. It is divided into six steps. The processing blocks are shown in Fig. on Data Mining Architecture(36).

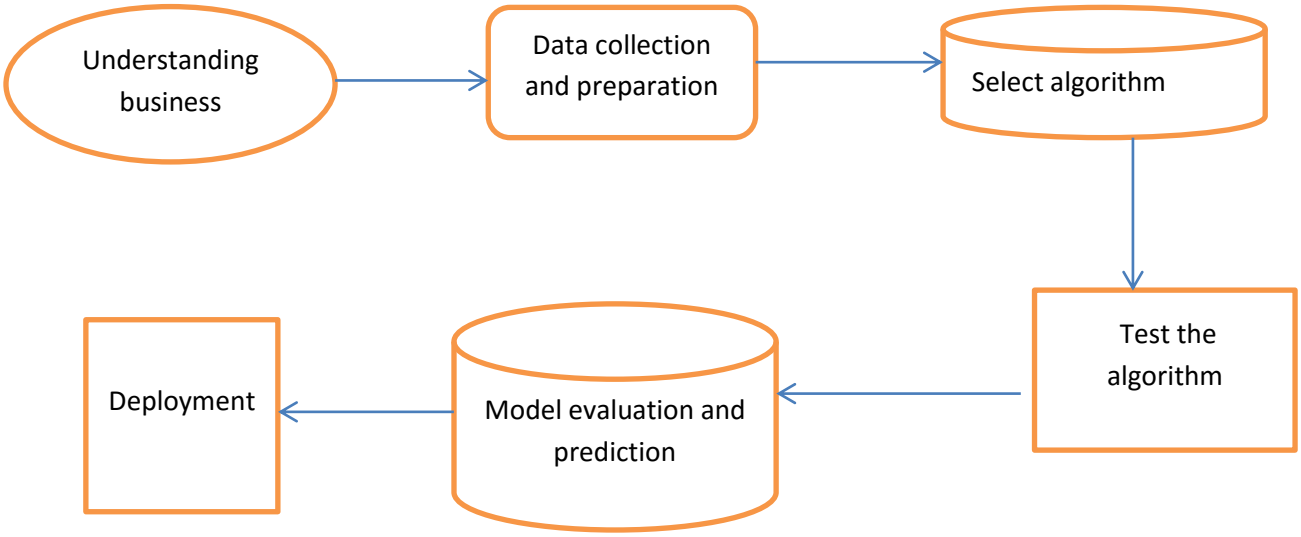

Figure 3.1 data mining architecture

Data collection: The first stage of the mining process is data collection from Hawasssa Adare general hospital southern nation nationality regional state. Data collection period from January to March 2017 by designed template adopted from WHO and Iran Ministry of health.

Data preparation: The data preparation stage is crucial for data analysis. Dataset stored in proper format were found to be insufficient. The WEKA Data Miner software requires

input to be provided in a particular format. Consequently, it was deemed necessary to convert the data **csv** format.

Data analysis: In the data analysis stage, data are analyzed to achieve the desired research project objectives. In the data mining techniques comprise a suite of algorithms such as decision tree and PART rule induction etc. In this study, we used classification technique that employed a j48 and PART algorithm.

Test the algorithm: At this stage, the desired algorithm and associated parameters have been chosen based on the parameter.

Knowledge evaluation and pattern prediction: This stage extracts new knowledge or patterns from the result dataset.

Deployment: The final stage of this process applies a previously selected model to new data to generate predictions.

3.3 Decision Tree Classifier

The use of decision trees is perhaps the easiest to understand and the most widely used method that falls into the category of supervised learning. Decision tree is powerful in its functions and a very popular tool for classification and making Prediction. The graphical representation of a simple decision tree using two attributes. A typical decision tree system adopts a top-down strategy in searching for a solution. It consists of nodes where predictor attributes are tested at each node, the algorithm examines all attributes and all values of each attribute although to determining the attribute and a value of the attribute that will "best" separate the data into more homogeneous sub-groups with respect to the target variable or class variable (42).

Decision trees are an approach of representing a sequence of rules that lead to a set or value. As a result, they are used for directed from mining, mainly classification. One of the main important of decision trees is that the model is quite reasonable since it precedes the form of generate rules that bases of the intervention. Another important classifier is easy and simple to implement. It doesn't have domain knowledge or additional parameter setting. It handle huge amount of dimensional data. It is more suitable for exploratory knowledge discovery. The results

attained from Decision Tree are easier to interpret and used for the prediction of trends and unknown patterns from the database. A decision tree is a classification tree when the outcomes are predicted interims of a class and a real number then it is known as regression analysis he leaves in the decision tree represent the class labels and combination of these class labels are represented by the branches(43).

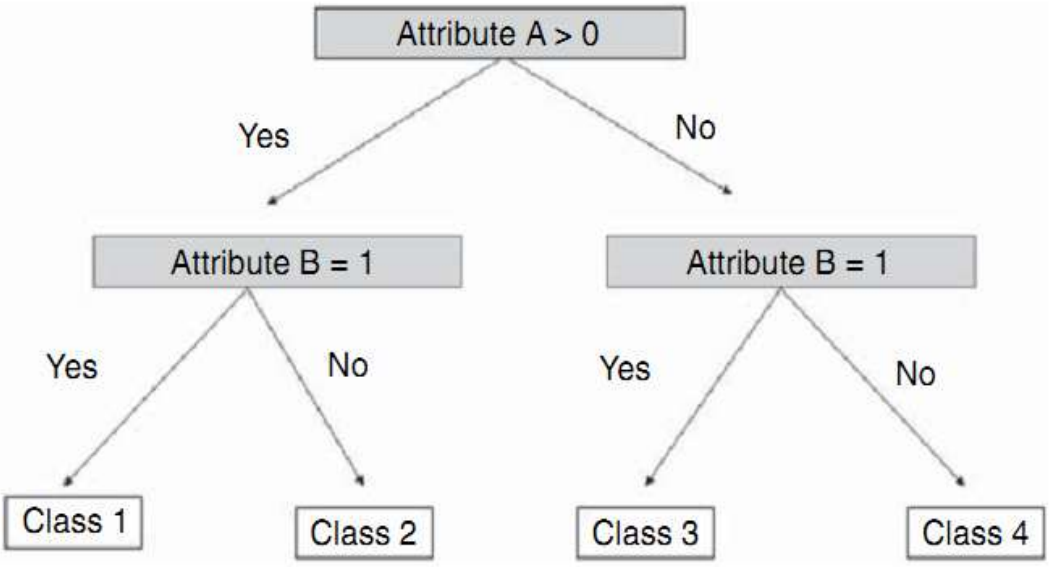

Figure 3.2 decision tree models

The decision tree algorithm used in this research is J48 algorithm, which implementation produces decision tree models. The algorithm uses the greedy technique to induce decision trees for classification. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. J48 generates decision tree; the nodes of tree evaluate the existence or significance of individual features.

Furthermore, a set of classification rules can be extracted from the decision tree by tracing the path from the root to each leaf (corresponding class). This set of rules can be consequently plugged into settle knowledge-based system. So, the researcher using J48 method in order to get the best fitted model that can appropriate to predict the pattern of diabetic screening identify the risk exposure consult patient to implement designed preventive methods in order to limit or delay the time of occurring diabetes(43).

3.4 Rule Induction

Rule induction is the process of extracting useful 'if then' rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. Knowledge represents IF-THEN rules for classification. An IF-THEN rule is an expression of the form Even though the pruned trees are more compact than the originals; they can still be very complex. Hence, generate rules to make a decision tree model more readable, it can be transformed into an IF-THEN decision rule. 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.

Rule induction or decision rule classifiers are set of IF-THEN classification. An IF-THEN rule induction 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(24).

3.5 Performance Evaluation

Once a predictive model is developed using the diabetes screening data, the model should be checked as to how it will perform for the future data that it has not seen during the model building process. The researcher has used three different classifiers to build the predictive model and in order to evaluate the performance of the model, for evaluation confusion matrix was used.

Confusion matrix is a useful tool for analyzing how well a classifier can recognize tuples of different classes. A confusion matrix is a table of size two by two. An entry, in the first rows and columns indicates the number of tuples of class one that were labeled by the classifier as class two. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, this explore in the following figure(44).

		True class	
		Positive	**Negative**
Predicted class	**Positive**	**true positive Count(TP)**	**False positive Count (FP)**
	Negative	**False negative Count (FN)**	**True negative Count (TN)**

Fig 3.2 simple confusion matrix

As shown in above figure a confusion matrix table of size two by two, the following measures can be calculated to measure the accuracy of the model, true positive rate, false positive rate, accuracy, Precision, recall, F – measure and ROC curve

The **true positive** rate of a classifier is estimated by dividing the correctly classified positives by the total positive count.

$$\text{True positive rate} = \frac{TP}{TP+FN}$$

The **false positive** rate of the classifier is estimated by dividing the incorrectly classified negatives by the total negatives.

$$\text{False positive rate} = \frac{TN}{TN+FP}$$

The **accuracy** of a classifier is estimated by dividing the total correctly classified positives and negatives instance by the total number of samples.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision is calculated by dividing correctly classified instances by the total number of correctly and incorrectly classified samples.

$$\text{Precision} = \frac{TP}{TP+FP}$$

F – Measure is calculated as the harmonic mean of recall and precision

$$\text{F- Measure} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

ROC (Receiver Operating Characteristics Analysis) curve

ROC analysis is performed by drawing curves in two dimensional spaces, with axes defined by the TP rate and FP rate. The TP Rate and FP Rate values of different classifiers on the same test set are often represented diagrammatically by a ROC Graph. On a ROC Graph, The value of FP Rate is plotted on the horizontal axis, with TP Rate plotted on the vertical axis. Each point on the graph can be written as a pair of values (x, y) indicating that the FP Rate has value x and the TP Rate has value y. The performance of different types of classifier with different parameters can be compared by inspecting their ROC curves.

CHAPTER FOUR

DATA UNDERSTANDING AND PREPARATION

4.1 Business Understanding

Understanding the business objective requirement from the business perspectives and then converting it in to a new knowledge by applying data mining gives problem definition and a design of preliminary plan to achieve the objective

4.1.1 Overview of Diabetes

The term diabetes mellitus describes a metabolic disorder of multiple etiologies characterized by chronic hyperglycemia with disturbances of carbohydrate, fat, and protein metabolism resulting from defects in insulin secretion, insulin action or both.

The cells of the body cannot metabolize sugar properly due to a total or relative lack of insulin. On other hand, the body breaks down its own fat, protein, and glycogen to produce sugar resulting in high sugar levels in the blood with excess byproducts called ketones being produced by the liver(45). It is classed as a metabolism disorder; Metabolism means the way our bodies use digested food for energy and growth. Disturbance of this process naming is diabetes pathology body can't utilize energy properly due to defects of pancreatic beta cell or shortage of production. Based on ethology diabetes are classified in to three main classes (45).

Type 1 diabetes incorporates diabetes that is basically an after effect of pancreatic beta cell devastation that prompts insufficient creation of insulin absolute insulin deficiency. This type can influence any age yet typically happens in youngsters and youth insulin dependent diabetes mellitus (IDDM). These diabetics can lead a typical life through regular insulin treatment, solid eating regimen, close checking and normal physical activity total case account 10% of all case. Characteristic of type1 diabetes not always present in other family members. Type I diabetes is not necessarily linked to obesity.

Type2 diabetes is the most well-known one that generally happens in adult peoples yet is progressively seen in kids and young people, as well. This type is called an insulin resistance in light of the fact that, in this type the body can create insulin yet possibly it is not adequate or the

body can't react to its impact leading glucose remains flowing in the blood. This type likewise incorporates LADA (Latent Autoimmune Diabetes in Adults), depicting a lessened number of individuals with diabetes type 2 who seem to have an insusceptible intervened loss of pancreatic beta cells . Numerous type 2 diabetics can control their blood glucose level through a healthy eating routine and an expanded physical movement. Their contributions are 90% of other case (46).

Gestational diabetes mellitus alludes to glucose intolerance with onset or first acknowledgment amid pregnancy because of ineffectively oversaw blood glucose. This collection must be nearly checked to control their BG level and minimize the danger for the child. This could be possible by healthy eating regimen, moderate physical activity and at times insulin treatment or oral drug. This contributes 10% of maternal mortality related to the pregnancy (47).

Symptoms may be absent when metabolic abnormalities are mild, but typical symptoms (i.e., thirst, polydipsia, polyuria, and weight loss) nausea, vomiting abdominal pain etc. and type 2 symptoms such as blurred vision, glycosuria etc. occur with the development of overt hyperglycemia. In severe cases, ketoacidosis or hyperglycemic-hyperosmolar state may occur, (48).

Usually type1 and sometimes type 2 diabetic patient come up with Acute complication severe forms ketoacidosis or a non-ketosis hyperosmolar state may develop and lead to stupor, comma and in absence of effective treatment, death. Most of the time type 2 patient manifest chronic complication or long-term effects of diabetes mellitus including progressive development of the specific complication of retinopathy with potential blindness, neuropathy that may lead to renal failure and with risk of foot ulcers, amputation, charcot joints and future of the autonomic dysfunction, sexual dysfunction and other central and peripheral cardiovascular disorder (49).

4.1.2 Risk Factor of Diabetes

Risk factors are conditions that increase the risk of developing diabetes disease. Diabetes risk factors broadly classify in to two: modifiable and non-modifiable risk factors. (49)**Modifiable Risk** factors means the patient can take measures/ intervention to avoid or minimize getting disease , which are Lifestyle modification reducing or managing the list of situation to control body weight by choosing dietary habits, body Mass Index (BMI) is keep normal <25kg/m2,

regular physical activity, manage/ control stress and condition that lead to stress, restrict alcohol uses, quitting smoking and other substances. Managing related conditions like high cholesterol and blood pressure. **Non-modifiable**, which means the individual or patient cannot change or avoid the risks that lead to diabetes. Non-modifiable risk factors are: age when age increase risk of diabetes increase, ethnic background, and family history of diabetes. IDF report the concordance of type 2 diabetes in identical twins is between 70- 90% inherit type 2 diabetes but both parents have type 2 diabetes the risk approaches 40% and history of comorbidity of other chronic disease like hypertension and other cardiovascular diseases(49).

4.1.3 Important of the Diabetes Screening

Diabetes screening is important mainly because of the following three reasons (23) According to IDF report in developing countries more than 83.8% Diabetes Mellitus patients are asymptomatic and unaware that they have the disorder, Type 2 DM may be present four up to a decade before diagnosis. As many as 50% of individuals with type 2 DM have one or more diabetes-specific complications at the time of their diagnosis

So that, screening is important because of high prevalence rate of diabetes disease and related complication. WHO technical committee report mentioned globally, 219 countries including Ethiopia 45.8%, or 174.8 million at age 20-79 years of all diabetes cases are estimated to be undiagnosed, ranging from 24.1% to 75.1% across data regions developed and developing country respectively. An estimated 83.8% of all cases of UDM (undiagnosed diabetes mellitus) are in low- and middle-income countries.(50)

In Ethiopia nationwide surveillance on occurrence of Diabetes Mellitus has not been made, some research conducted but well not documented and update regularly. Wolde.M reviewed in Jimma, Addis Ababa, Gonder city and surrounding district even if this paper is done in diabetes related complication in Ethiopia. That does not show the prevalence of diabetes in the country see the table 4.1.

The researcher indicates that major DM related complications include: hypertension, 12.1 % (1976-1997) to 34.1%, 2005 to 2009 neuropathy, 27.7%, in 1976-1997 to 29.5%, and DM foot ulcer disease1.7% (1976-1997) to 4.6 %. The prevalence of these diseases has increase. So, the

paper shows indirectly the prevalence diabetes increase through time. The following table demonstrate diabetes related complication(51, 52).

Table 4.1 tabular presentation of diabetes complication in selected Ethiopia (M, Wolde 2013)

Distribution of Diabetes Mellitus association complications on selected studies conducted in Ethiopia								
province	year of study	Sample size	Hypertension %	Retinopathy % (eye disease)	Neuropathy	Nephropathy (renal disease)	foot disease	reference
Addis Ababa	1976-1983	849	12.1	33.3	27.7	27.7	0	10
Addis Ababa	1996-1997	283	18.4	31.4	35.2	23.3	1.7	11
Addis Ababa	2005-2009	724	34	15.5	12.4	32	0	13
Addis Ababa	2005	229	34.1	33.2	10.5	21	0	14
Jimma	2008	305	24.9	33.8	29.5	15.7	4.6	15

The other major activities for prevention is screening which aim at early detection of diabetes and prompt effective management of the condition with the propose of reversing condition and /or halting its propagation. The strategy aimed at the detection of as yet undiagnosed cases of diabetes, aging activities can be targeted at population or high risk groups or at risk individual and prevent complication and disability due to diabetes.

Early detection of diabetes risk factors could give great value to reduce the prevalence of diabetes by delaying or preventing diabetes related complication of all people (53). Another benefit of early detection is also important for focusing of life style and behavioral modification that pushes to change the government strategies and implementation for health care planer and police makers(48).

Screening strategy depends on the underlying prevalence of diabetes, structure of the local health-care system, and the economic condition of the country. Screening for diabetes is not an easy task; it needs an analytical, organizational, and financial challenge. The organizational and financial aspects are the biggest limiting factors to implement the program. Several strategies

have been suggested and proposed for facility based screening. If possible to cop up within the local health-care system or integrated with existing health care delivery program. Some of the strategy support to deliver appropriate follow-up, from screening disease, improving care and follow up the patient and reminding patient before getting disease.

The common and best methods for screening diabetes prevalence and incidence is performed by fasting blood glucose (FPG), this is not feasible to implements for screening because most patient visit health facility after taking meals and can't be afford the fee for laboratory examination .

The second method for screening is Selective screening method; this method is done by distributing written questionnaire or verbal interview to the eligible population. This questionnaire should identify those individuals who are at high risk of having diabetes in the community should be referred to a physician for consideration of diagnosis.

This screening method is also difficult to implement in community and facility level. Because majority of the patient is not literate and in each visit doing the same procedure that make poor quality due to the burden for health care workers. This method also requires high cost.

The final method is opportunistic screening strategies; this strategy is used for the detection of people with diabetes contact of health service for any other reason during physical and laboratory examination in health institution.

This method is the relatively best for identifying pre diabetes patient with the minimum cost and reducing the screening time. And this method increase the incidence rate of diabetes patient rather than the previous two strategies and handling of patients entered the screening program continued follow up every three years at the age 30 and above(54).

The outcome of screening program through above listed strategy selected based on feasibility, in terms of cost and integration with existing local health program to the detection of asymptomatic individuals who are likely to have diabetes. To this end variables such as epidemiological or environmental factors, such as background of patient, age, sex , dietary habits of individual physical activities, genetic inherit of patient previews diabetes history of parents, clinical or diabetes related sign and symptoms like polyphagia, polyuria…etc. Life style, behavioral factors

and other history of comorbidity factors are also taken in this project for data collection and achieve the objective of data mining and construction of predictive model to support screening program.

4.2. DATA UNDERSTANDING

4.2.1. Data Sources Description

Source of data used in this study is Hospital based patient data conducted in for three successive months in the Adare hospital, Hawassa, Ethiopia. The main reason to select the Hospital is, there are concerning factors such as diabetes patient flow or data yield, new HMIS implementation, availability of organized diabetes and chronic patient department in the hospitals. All hospitals on city including research area there are no automated data management system available. The researcher select relevant attribute in consultation with public health expert and clinician who work in hospital and interview patient by designed templet. Hence, the researcher first encoded all the data in an Excel format. After the data was encoded, the entire dataset is put in one file having many records. Each record corresponds to most relevant information of one patient. Next, Pre-processing techniques was applied to make it appropriate for mining purpose.

4.2.2. Data Understanding

This step is crucial for understanding core business perspective and identifying the required attribute in order to meet objective of study. In addition to this they are familiar to all dada element relevancy and validity of data set. On the other hand, understanding the characteristics of data after completing the process selected relevant attributes which address pre diabetes or undiagnosed patient screening by applying data mining method.

Demographic characteristics include age, gender, marital status, and level of education. A family history of diabetes was defined as a person that have a diabetic parents or ancestors.

In this study anthropometric measurements were taken to identify the BMI and lifestyle risk factors include the following variables: cigarette smoking; alcohol consumption, asses the patient stress and associative risk factor. Measuring stress is very difficult because it is subjective but the study focuses on job related stresses. Therefore, the researcher took type of job engagement as an

indicator, so the variables classified as governmental, non-governmental and NGO company, self-job and retired(55).

Table: 4.2 List of Variables in the Initial Dataset

No	Attribute name	Data type	Description
1	age	Numeric	Age of the patient in years
2	Sex	nominal	Sex of the patient, male or female
3	Residence area	nominal	Residence of patient, either urban or rural
4	Job status	nominal	Job status of patient, self-employment, governmental, non-governmental, retired
5	BMI	Numeric	Body mass index of patient calculated by weight KG in divided by height square kg/m2
6	History of physical activities	nominal	History of patient regular physical activity,
7	Category of food usually eat	nominal	Type of food usually eat , carbohydrate, protein and fat
8	History of hypertension	nominal	History of hypertension of patient, yes or no
9	Family history of diabetes	nominal	History of family having diabetes, yes or no
10	History of smoking	nominal	Smoking status of patient, yes or no
11	History of alcoholic	nominal	History of alcohols consumption of patients, yes or no
12	History of CVD	nominal	History of heart failure of patients ,Yes or no
13	Polyuria	Nominal	Symptom of excess urine of the patient yes or no
14	Polyphagia	Nominal	Patient manifest Excessive appetite of food, yes or no
15	Polydipsia	Nominal	Excessive or constant thirst occasioned by patient, yes or no

16	Visual complaints	Nominal	Sing of blurring of vision of patient yes or no
17	Diabetes status	Nominal	The final outcome of patients , diabetes or not diabetes

4.2.2.1 Data Processing

Data processing techniques is one of the crucial steps in data mining. Data collected from any sources does not normally show problems at the first glance. So applying consecutive steps is mandatory, so as to solves data quality problem and designed a proper techniques for handling the problem. Applying the steps before mining can substantially improve the overall quality of patterns mined and/or the time required for the actual mining. Also it is an important step in the knowledge discovery process, because quality decision needs a quality data(56).

4.2.1.2 Data Field Selection

This step is important for selecting the relevant attribute of data set for mining. In case of this research project there is no well-organized data base used for mining. Patient data is available in the form of hard copy format. So the investigator developed template by adopting from different literatures like IDF, WHO and Iran ministry of health screening tools and some modification was done to make it appropriate to the study. Then all patient data encoded by developed template. Using this method reduces redundancy and irrelevant attributes through this process. Some attribute omitted before actual data collection like *name of patients*, *address of patient worda*, kebele, *house no*, *patient medical record number* and *type of visit either new or repeated* before actual data collection process.

4.2.1.3. Data Cleaning

Real-world data tend to be incomplete, noisy, and full of inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data(24). In case of this project there is no data discrepancy because there is no designed database in the diabetic clinic. So, data collected through the designed template but some minor problems were happened during data collection like missing value, misplaced and type error during complete designed tools.

Missing value: - In the collected data for this research work there were some missing values like food usually used, age of patient. This is corrected by the time of next visit and some of them with the help of the domain experts special types of diabetes based on characterization risk factor of patient so that all the missing values were filled with the appropriate value.

Detecting noisy data and outliers: Noise is a random error or variance in a measured variable. A database may contain data objects that do not comply with the general behavior or model of the data these objects are outliers. In this research age type single digits below target age group, corrected by investigator and consult senior expert who charge on clinic(24).

4.2.1.4 Data Discretization

In some of data needs discretize to smoothing for mining. Reduce data size by dividing the range of a continuous attribute into intervals against to the standard of WHO. Interval labels can be used to replace actual data values. For example, smoothing techniques including binning, and dividing value by hierarchal derived new attribute construction are the most used ones. From the dataset the "AGE" and "BMI" attributes are continues value change to discrete value thought discretized (binned) process.

Table 4.3 Summary of Derived Attributed with Their Values

Sr no	Original Attributes	Existing value	New value
1	Age of participant	Age, cat	20-30,31-40, 41-50,51-60,61-80, >80
2	Body mass index(BMI)	BMI cat	W<=18.4 under weight, X=**18.5-25kg/m2 = Normal Y=**26-30 g/m2= overweight,30-40 kg/m2 = obese **Z >=** 40 kg/m2 = very obese

4.2.1.5. Final Selected Data Set

So far different corrective measures were taken on the remaining attributed. After finishing the data cleaning process, we saved the file into csv format prepare to mining software WEKA. Attributed such as category under background *age, sex, educational status, Occupation* and the

environmental factor *regular physical exercise*, *BMI*, *type of nutrition uses*, genetic predisposing factor *family history of diabetes*, other factor comorbidity *hypertension* and *history of CVD* and *symptoms classified Polyuria, Polyphagia, Polydipsia and Visual complaints* are selected as final summary of the dataset constructed ready for experiments with the use of algorithms is depicted in table 4.4 as follows.

Table 4.4 Summary of the selected dataset ready for mining

Sr.no	Categories	Final result
1	Number of attributes(finally clean)	16
2	Number of instances	4529
3	Number of classes attribute	1

CHAPTER FIVE

EXPERIMENTATION AND ANALYSIS

This chapter built experiments carried out to construct predictive model together with their analysis in order to generate the best algorism. The experiments were run in WEKA software 3.6.9 prepared dataset. This addressed the objectives of the research with respect to the minimum error that consists of 15 attributes and 1 class attributes. PART rule induction and J48 decision tree algorithms are used for constructing predictive model.

A dataset is imbalanced if the classification categories are not approximately equally represented the status of patient. Performance of imbalance data is deviates to doming value and bias the predictive accuracy. In the case of pre diabetes screening data the class variable has a higher imbalance with ratio 1:4 Therefore; the investigator decide to balance reduce the class attribute 1:1 ratio with minority classes to prevent false positive result.

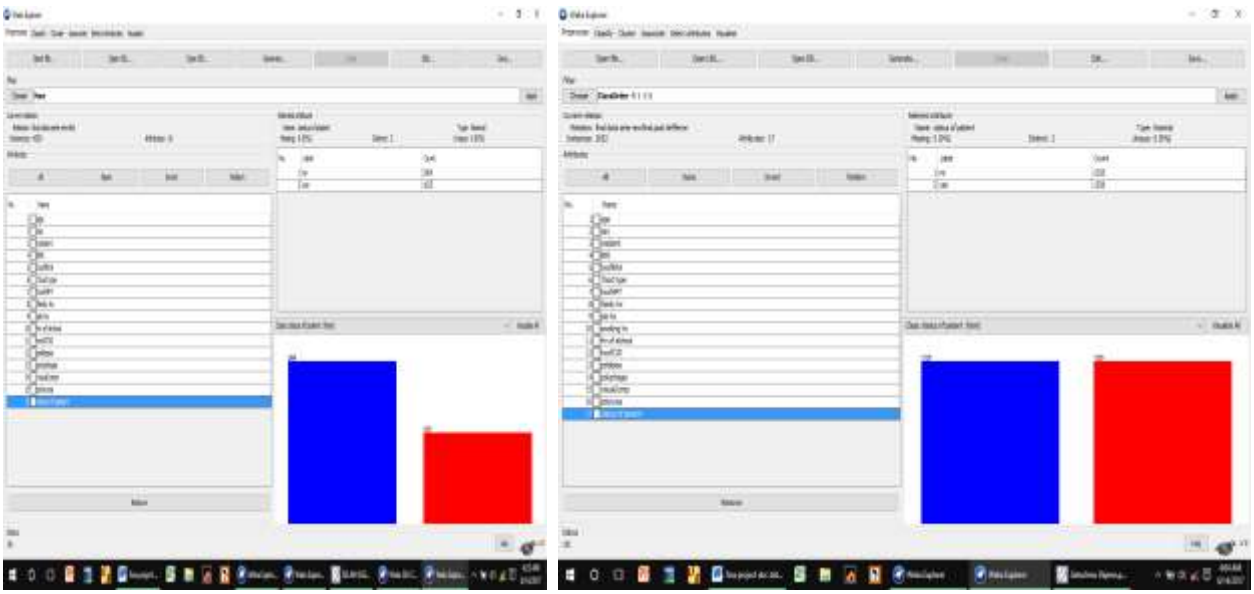

Figure 5.1 Side by side review of the class attribute in diabetes and pre diabetes patient (left side) Original data; (right side) balanced data.

Figure 5.1 shows a side by side review of the class attribute is status of patient reduce the majority class. Originally there were 3204 records in the majority class and only 1325 records in the minority class but after applying balance by reduction the class is equal with the minority class.

5.1 Model Building

In this study was made to develop a model that enables to predict the pattern of the identify risk factor of diabetes. For creating predictive model a total instants of 4,529 records were used for training and testing. The validations were 70% split for training and 30% for test is used for learning the parameters of the model in order to produce hypotheses and evaluates the accuracy of the model/hypothesis in predicting. Effectiveness of the predictive algorithms is checked by using 10-fold cross validation. This performance estimation approach has been proved to be statistically good enough in evaluating the performance of data mining classifier algorithms. Overall classification accuracy was examined using the confusion matrix computing , TPR rate, FPR rate, precision F-measure, and ROC area were used to evaluate and compare the performance of the models(57).

In classification process outcome is predicted from a given input or future attributes relation to class attributes. For this purpose the algorithm process a training set which consists of an attribute, check the algorism feat with given data set and the outcome is called prediction attribute. The algorithm progresses by finding the relationships between the input attribute set i.e. the training set. The input is then analyzed to produce a prediction and how good an algorithm is depends upon the predictions made by the algorithm. Prediction rules are used for knowledge expression in the form of IF-THEN rules, IF part is known as antecedent which consists of a conjunction of conditions and the THEN part is known as consequent which gives the prediction whether satisfies the antecedent or not(58).

The experiment used for achieved this project all 17 attributes and the selected by used best of first 9 attributes includes class presented the following table.

Table 5.1 list of attributes in all and selected attribute used by best of first

	All attribute (17)	Selected attribute (10)(used best first)
	1. Age 2. Sex 3. Resident 4. BMI 5. hxofRFA 6. Food type 7. hxofHPT 8. Family history 9. Job type 10.Smoking history 11. Hx of alcohol 12.HxofCVD 13.Polydipsia 14.Polyphagia 15.Visual/comp 16.Polyurea 17.Status of patient	1. resident 2. food type 3. hxofHPT 4. Hx smoking 5. Hx of alcohol 6. Hx of CVD 7. polydipsia 8. polyphagia 9. visual/comp 10. status of patient

5.1.1 Model Building Using J48 Classifier with all attributes

In the experiment is conducted decision tree J48 algorithm with pruned and un-pruned and all 16 attribute and 1 class. Attribute selection best first methods by the weka machine selected 9 attributes and one class. The performance of J48 classifier in predicting diabetes screening evaluated present in the following experiment.

Experiment1

For this scenario models were built used J48 un-pruned with all 16 attributes and 1 class the results of the experiment is presented in experiment below.

```
Correctly Classified Instances        2316               87.3303 %
Incorrectly Classified Instances       336               12.6697 %

Total Number of Instances              2652

=== Detailed Accuracy by Class ===
```

```
Class     TP Rate   FP Rate   Precision   Recall F-Measure   ROC Area
          0.895     0.149      0.858     0.895     0.876      0.912    no
          0.851     0.105      0.89      0.851     0.87       0.912    yes
.         0.873     0.127      0.874     0.873     0.873      0.912

=== Confusion Matrix ===

   A    B   <-- classified as
 1187 139 |    a = no
  197 1129 |    b = yes
```

This experiment of un-pruned J48 decision tree experimental train and test 70% of the records for training and 30%, of the instants for test and 10 fold cross validation the results of correctly classified, accuracy of model, true positive and false positive rate Furthermore the result obtained from these experiments is summarized in table 5.2

Table 5.2 Summary of the four Decision Tree Experiment Results

Algorithm	No of attribute	True positive		Rock area		Accuracy of model	
		10 fold	70% split	10 fold	70% split	10 fold	70% split
J48 all pruned	17	83.7%	80.2%	86.4%	85.3	82.1%	80.2%
J48 all un-pruned	17	81.6%	80%	82.5%	80.8%	80.57%	82.88%
J48 selected pruned	10	82.6%	83.5%	84.1%	83.5%	84.1	80.6%
J48 selected un-pruned	10	83%	87.5%	83%	85.9%	81.67%	83%

The tested model has accuracy of pruned and un-pruned all 17 attribute 80.57% using 10 fold cross validation and 82.88% accuracy using 70% split test options. Moreover the model has a true positive rate of 81.6%and rock area of 82.5%for 10 fold cross validation and 80.8% rock area for 70% split test. In the meantime the second experiment used 10 attribute that experiment one identified to be statistically significant to construct the decision tree. It exhibited the same accuracy, true positive rate and true negative rate as to that of experiment one.

The best decision tree model produced was from j48 all attribute un-pruned. The model shows a better performance evaluation than other models. The 70% split test model also scored a better performance than 10 fold cross validation.

Therefore, the all 17 attributes, used to build the decision tree for j48 with 70% split test option, are taken to be **82.88%** statistically significant in splitting the decision tree. Furthermore, opinions gathered from the domain experts from working clinician, these indicated that attributes have a great role in the prediction diabetes.

5.1.1.1 Confusion matrix for J48 decision tree classifier

The confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes. The confusion matrix for the decision tree shown in table 5.2 demonstrate that out of the total 2652 records 1115 records are correctly classified as category "yes" and 1076 records are correctly classified as category "no". The classifier incorrectly classified 211 records categorize as "no" and 250 records categorize as "yes". It has 361 attributes are misclassified both category of 'yes' or 'no'. While the accuracy of the classifier to correctly predict the class value as "yes" and no is 82.88% which is better algorithm from the above.

Table 5.3 Confusion Matrix for J48 Decision Tree Model

Confusion Matrix		
A	B	Classified as
1115	211	A= yes
250	1076	B= no

5.1.1.2 ROC Analysis for J48 Decision Tree Model

ROC analysis provides tools to select the best models and to discard suboptimal ones. ROC analysis is related in a street way to cost/benefit analysis of diagnostic decision making. Figure 5.2 shows the area under ROC for the pre diabetes screening Instances. Class value yes gives the ROC accuracy of 82.88% of algorithms selected from all 17 attribute with un-pruned experiment.

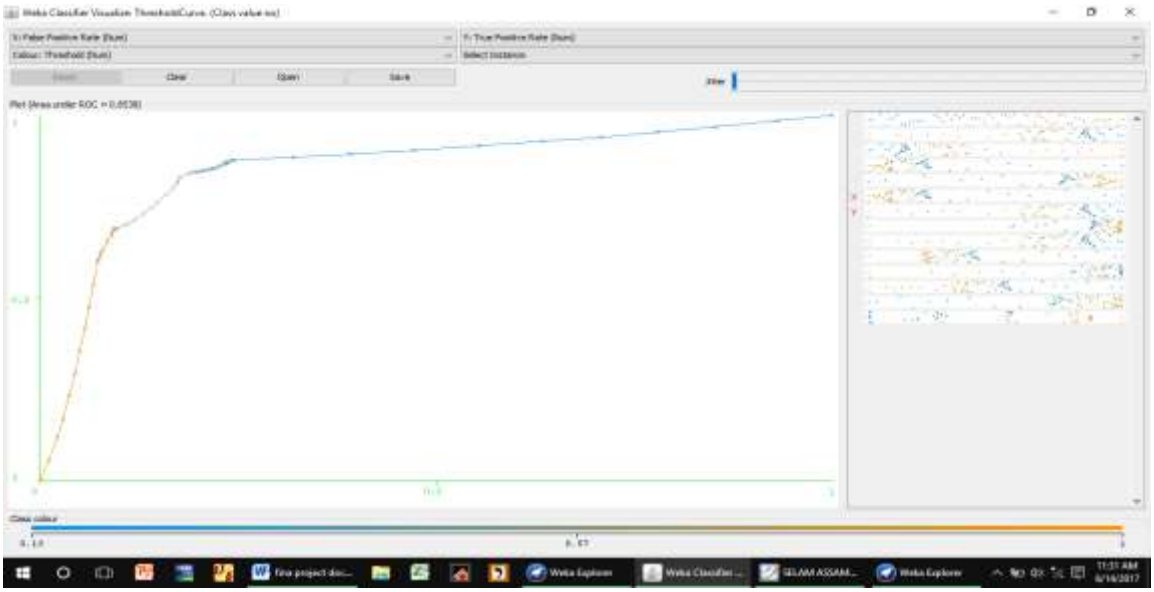

Figure 5.2 ROC curve of the decision tree model

5.2 Experimentation with PART Algorithm with all attributes

PART rules induction algorithm extracts the given attributes from final dataset of the project. The detailed procedure of the algorithm extracting rules is explained in by the data mining methodology. The algorithm builds partial decision trees and reads a path from the root of the tree. PART has almost a similar set of parameters with J48 algorithm that can be adjusted to build better model from datasets. Like the J48 experiments, PART experiments were also performed in two experimental settings based on the parameter pruned and the un-pruned algorithm. The experimental settings are indicated based on the next scenario i.e. PART Experiment with pruned and Experiment un-pruned with all attributes and selected 10 attribute.

In the given experiment, the value of the performance measure is partially difference observed among the model. So that makes comparison necessary to select the one with other relatively better measures of performance. The tool used to measures Performance is based on confusion matrix, such as TPR, false positive rate, Precisions ROC and accuracy classified of given instants the sample experiment results demonstrate in the following.

```
Instances:    2652
Attributes:   17
=== Summary ===

Correctly Classified Instances        2571               96.9457 %
Incorrectly Classified Instances        81                3.0543 %
Total Number of Instances             2652
=== Detailed Accuracy by Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
Class
                 0.98      0.041      0.959     0.98      0.97       0.998 no
                 0.959     0.02       0.98      0.959     0.969      0.998 yes
Weighted         0.969     0.031      0.97      0.969     0.969      0.998

=== Confusion Matrix ===

    A    B   <-- classified as
 1300   26 |    A = no
   55 1271 |    B = yes
```

The output of PART algorism, the tested algorithm is 'true' un-pruned experiment PART has accuracy or correctly classified 1300 instants, 1271 correctly classified under category of false instances. The experiment were classified 70% of the records for training and 30%, of the instants for test and 10 fold cross validation the output algorithm correctly classified, accuracy of model, true positive and false positive rate tools for accuracy. The results of the experiment in presented as following table.

Table 5.4 Summary of the four PART rule induction Experiment Results

Algorithm	No of attribute	True positive		Rock area		Accuracy of model	
		10 fold	70% split	10 fold	70% split	10 fold	70% split
PART all pruned	17	87.2%	91.4%	84.6%	95.8%	84.01%	90.7%
PART all un-pruned	17	80.9%	98%	86.7%	99.8%	87.07%	96.9%
PART selected pruned	10	83%	82.5%	87.3	87%	81.3%	80.42%
PART selected un-pruned	10	82%	85.4%	86.9%	89.2%	80.65%	83.74%

5.2.1 Confusion matrix for PART rule induction classifier

The PART classifier one of selected algorithms test diabetes screening project. The confusion matrix for the PART classifier shown in table 5.4 demonstrate the total 2652 records 1300 records are correctly classified as category "yes" and 1271 records are correctly classified as category "no". The classifier incorrectly classified 26 records categorize as "no" and 55 records categorize as "yes". It has totally 81 attributes are misclassified both category of 'yes' or 'no'. While the accuracy of the classifier is correctly predict the class value as "yes" and no is 96.9% which records PART classifier is better result than that of the decision tree j48 algorithm.

Table 5.5 Confusion Matrix for PART rule induction Model

Confusion Matrix		
A	B	Classified as
1300	26	A= yes
55	1271	B= no

5.2.2 ROC Analysis for J48 Decision Tree Model

ROC area analysis power full tools to select the best models. ROC analysis is related in a directly measure the cost/benefit analysis of diagnostic decision making. Figure 5.5 shows the area under ROC for the pre diabetes screening Instances. Class value yes gives the ROC accuracy of 99.79% of algorithms selected from all 17 attribute with un-pruned experiment is .

Figure 5.3 ROC curve of the PART rule induction model

The model built by Binary both decision tree and PART rule induction have four scenarios like: all attributes without pruning, all attributes with pruning, some selected best first attributes without pruning, some selected by best first attributes with pruning.

The PART experiment tested algorithm result in above table has accuracy of pruned and un-pruned all 17 attribute 96.9% using 10 fold cross validation and 87.03% accuracy using 70% split test options. Moreover the model has a true positive rate of 80.9%and rock area of 86.7%for 10 fold cross validation and 98% rock area for 99.8% split test. In the same with first scenario, the second experiment used 70% attribute that experiment one identified to be statistically significant to construct the rule induction. It exhibited the same accuracy, true positive rate and true negative rate as to that of experiment with the decision tree.

5.3 Model Evaluation

The model selection criteria were done based on the statistical summary obtained from the WEKA machine learning open source software environmental for knowledge analysis. Based on the evaluation parameters are tested algorithm to compare classifiers done using the two mining algorithms; accuracy of the model, sensitivity (TPR), False Positive Rate (FPR) and area under the ROC curve. The overall performance of each of algorithm the best two classifier models were all 17 attribute and selected 10 attribute including class with pruned (false) and un-pruned (true).

In this project work are presented number of relevant experiments had been carried out with J48 decision tree and PART rule induction classification algorithms identify as best performer, i.e. J48 decision tree algorithm, and the PART rules algorithm to build a better predictive result that that extracted form the best scored accuracy that of all experiment. Through this process identify the best algorithms that can predict early diabetes screenings integrate with regular OPD and other chronic patient follow unit. From the tested experiments all 17 attributes with un-pruned PART rule induction were identified best accuracy.

5.3.1 First Scenario

The first experiment was to evaluate the performance of J48 classifier algorithm by considering two cases: building two models using the given dataset with all 17 attributes and the selected 10 attributes. The goal is to investigator whether attribute reduction improves or degrades the

performance of the model built using the 70% training set and the remaining 30% test set and 10 fold cross validation. Considering the above statistical values labeled under each of the evaluation parameters the result 82.88% accuracy by j48 algorithm. Model was selected for next comparison with results of PART rule induction.

5.3.2 Second Scenario

The experiment was evaluated the performance of PART rule induction classifier algorithm is also considering two cases: on the top of this scenarios were considered to do the experiment: building model using pruned and un pruned with whole17 attribute and 10 attribute are selected '*weka attribute selection best first*' the experiment was analysis and conducted both all attribute and the selected attributes.

The PART rule induction experiment both un-pruned and pruned experiment tested model has discovered a better performance algorism. There for 17 attribute test pruned and un pruned at the same time selected attribute selected 10 attribute test pruned 'false' and un pruned 'true' algorism for this scenario the second experiment PART rule induction with un pruned all 17 attribute splitting 70% training and the remaining 30% test is best algorism with accuracy 96.9%, TPR 98% and ROC 99.8% model was constructed the rules used for the purpose of design prototype.

To end with the project the rules PART un-pruned with all attribute is best attribute accuracy 96.9% and J48 with whole attribute model the second best performance accuracy 82.88% but, the difference of accuracy and correctly classified of instances. From the above two algorisms are compared based on parameters. PART rules induction high accuracy than j48. The investigator extract rules for this project applying for designing pre diabetes screening model(59).

5.4 Rule Generated from the Selected Model

Based on experiment PART rules with all attribute un-pruned are high accuracy finally generated rule learner with the specified scheme has results PART total of 325 rules. Out of these the rules which are highly predictive are selected 25 consider the mandatory attribute by investigator based on the finding and relevant to the domain knowledge or area of specialized people. The following are selected best rules generated and input for model building prototype for screening.

Family or parent history of diabetes

As literature of American diabetes association have stated the relationship between hereditary or parent to children the incidence is 40-70% which has not diabetes. The risk depends on ancestors, immediate parents and twins. From Rules generated related to family history consulted with respective clinician the following rules are pertinent to related with titles are selected for building model.

Rule 1:
IF Family history of diabetes = yes AND sex = F AND food type = fat AND Hx of Hypertension = yes AND Hx of Regular physical activities = yes: THEN Yes (3.0/1.0)
Rule 2:
IF Family history of diabetes = yes AND Food type usually used = fat AND job type = self AND age = 40-50yeas: THEN Yes (5.0)
Rule 3:
IF Family history of diabetes = no visual/comp = yes AND History of alcohol = no AND polydipsia = no: AND BMI = 25-30kg/m2 THEN Yes (4.0)
Rule 4:
IF Family history of diabetes = no AND polydipsia = yes AND Hx of Hypertension = no AND age = 30-40 years AND Resident = Urban AND Polyurea = no: THEN yes (5.0)
Rule 5:
IF Family history of diabetes = no AND History of alcohol = yes AND History of cardiovascular disease = no AND BMI = 30-40kg/m2 AND age = 40-50 Years: THEN Yes (3.0/1.0)

Rule selected by age and association risk factor

Different literature and clinician agree age is non-modifiable risk factor is getting disease. Age and diabetes proportion mention in above age of >55 twice increase the susceptible than that of age less than 45. The hospital trained report trends becoming high. Health care professional and study of scholar even if age non- modifiable risk factor but delaying or extended the risk having disease by intervention act on indirectly association with age like limited the consumption of alcohol, regular physical activities, manage the body weight. Rules of generated PART rule induction selected for model building discussion with the domain expert the following.

Rule 1:
IF Age = 30-40 Years AND Polyphagia = yes AND sex = F AND Visual/comp = no AND Polydipsia = yes AND Family history of diabetes = no AND BMI = 18-25kg/m2 AND History of cardiovascular disease = no AND Job type = self: THEN Yes (6.0/2.0)

Rule 2:

IF Age = D: AND Polyphagia = yes AND Sex = F AND Polyurea = no AND History of Smoking = no AND Resident = Urban AND job type = Government THEN Yes (7.0)

Rule 3:

IF Age = 30-40 Years AND Polyphagia = yes AND History of alcohol = yes AND History of cardiovascular disease = no AND BMI = 18-25kg/m2: THEN Yes (9.0)

Rule 4:

IF age = B: AND polydipsia = yes AND Polyphagia = yes AND History of Smoking = no AND Hx of Hypertension = no AND Food type = fat AND History of cardiovascular disease = no THEN Yes (7.0)

Rule 5:

IF Age = 50-60 years: AND Polydipsia = yes AND polyphagia = yes AND History of Smoking = no AND Family history of diabetes = no AND Food type = Protein AND History of cardiovascular disease = no AND History of alcohol = no THEN Yes (10.0)

Rules related to BMI

Obesity is major risk factor for developing non-insulin dependent diabetes and the other chronic non-communicable disease like developing coronary heart disease, hypertension and an increased this contribute majority of mortality in both developed and developing country. But BMI attribute has not been selected by Best first machine learning, but rules induction extracted from all 17 attributes, the accuracy of algorithm is 96.9%. Sample of rule generated by the algorism to build predictive model is listed below.

Rule 1:
IF BMI = 25-30kg/m2 AND Polyphagia = yes AND polydipsia = yes AND History of cardiovascular disease = yes AND Hx of Hypertension = yes: THEN Yes (6.0)

Rule 2:

IF BMI = 25-30kg/m2 AND History of alcohol = no AND visual/comp = no AND Food type = protein AND Hx of Hypertension = yes AND Regular physical Activity = no AND age = 40 50years AND polydipsia = no: THEN yes (3.0.1)

Rule 3:

IF History of cardiovascular disease = no AND job type = self AND Food type = Carbohydrate AND BMI = 18-28kg/m2: THEN Yes (3.0)

Rule 4:

IF Hx History of alcohol = no AND History of cardiovascular disease = no AND age = 60-70years AND BMI = 18-25kg/m2 AND sex = F: THEN Yes (6.0/3.0

Rule 5:

IF BMI = K AND Hx History of alcohol = no AND History of cardiovascular disease = no AND visual/comp = no AND family history of diabetes = yes AND resident = rural: THEN Yes (2.0)

Rules generated by attribute of regular physical activities

Many of the risk factors for diabetes, heart disease, cancer and pulmonary diseases are related to lifestyle the researcher suggested to prevent chronic non-infectious disease is regular physical activities. Physical inactivity reduces approximately 80% of non-insulin dependent diabetes. Patient associated with poor history of physical activity early having diabetes this is true for other non–communicable. The weka software was also selected by the Best first as the significant with eight attribute some of the Rule extract from physical activity is selected and listed below.

Rule 1:

IF History of regular physical activity = yes AND History of alcohol = no AND History of cardiovascular disease = no AND age = 30-40years AND BMI = 25-30: THEN Yes (5.0)

Rule 2:

IF History of regular physical activity = yes AND Polydipsia = no AND History of cardiovascular disease = no AND Job type = Non-governmental AND polyurea = yes: THEN Yes (4.0)

Rule 3:

IF History of regular physical activity = no AND polydipsia = yes AND Polyurea = no AND Food type = protein AND Age = 30-40years: THEN Yes (7.0)

Rule 4:

IF History of regular physical activity = no AND History of cardiovascular disease = yes AND Age = D AND resident = rural: THEN Yes (3.0)

Rule 5:

IF Family history of diabetes = yes AND Food type = protein AND History of regular physical activity = yes AND Age = 20-30years AND Resident = Urban: THEN no (8.0)

Rule generated comorbidity

The New England journal of medicine published the history of hypertension from parental history of diabetes were significant predictor of the diabetes. Chronic disease or comorbidity deteriorate the immunity of the individual and exposed to the other opportunistic infections and non- infectious disease. Diabetes is also one of the immune compromise natures so getting other chronic disease possibility is high. Mining software of the Best first has not been selected 10 attributes. From high accuracy algorithm, some rules are selected with consultation domain expert and display the following.

Rule 1:

IF History of cardiovascular disease = no AND polydipsia = no AND history regular physical activity = yes AND BMI = 18-25kg/m2 AND History of alcohol = no AND History of hypertension = no AND Food type = protein: THEN YES (9.0/3.0)

Rule 2:

IF History of cardiovascular disease = no AND History of hypertension = yes AND Food type = carbohydrate AND history of regular physical activity = no AND history of alcohol = no AND job type = retired: THEN NO (20.0/3.0)

Rule 3:

IF History of cardiovascular disease = no AND polydipsia = no AND history of regular physical activities = yes AND BMI = 18-25kg/m2 AND history of alcohol = no AND History of hypertension = yes: THEN YES 7.0/1.0)

Rule 4:

4. History of cardiovascular disease = no AND polydipsia = no AND History of regular physical activity = yes AND BMI = 18-25kg/m2 AND history of alcohol = no AND history of hypertension = yes: THEN YES (7.0/1.0)

Rule 5:

IF History of cardiovascular disease = yes AND History regular physical activity = no AND polydipsia = no AND Food type = protein AND Age = 40-50years: THEN YES (4.0)

This all listed rule derived from the PART rule induction algorism for the purpose of demonstration and codding on the prototype even though total number of rule greater than 500. From listed some of the pertinent rules some of them are discussed on next topic 5.10.

5.5 Prototype Development

A typical decision support system consists of five components: The screening model consists, the data base which store user information user name and pass word, rules, the user interface, and the users. One of the major differences between decision support systems employing data mining tools and those that employ rule-based expert systems rests in the knowledge engine(53).

In the decision support systems that utilize rule-based expert systems, the inference engine must be supplied with the facts and the rules associated with them that are often expressed in sets of "if–then" rules. In this sense, the decision support system requires an extracted knowledge on the part of the decision maker in order to provide the right answers to well-formed questions. On the contrary, the decision support systems employing data mining tools on the part of the decision maker. Instead of the system is designed to find new and unsuspected patterns and relationships in a given set of data. This is assisting health care professional, reduce time of decision making simple and easy to implement and integrated to other service delivery system and the management of health care planning.

5.5.1 Diabetes Screening CDS's User Interface

The user interface is a channel for communication between the system and the end-user. Therefore, in order to design the CDS to be an interactive tool, decision was made by referring the set of rules or command in a simple manner. Examples of information to be shown are the

consequences made by against with the set of rules. The events made back of screen and an explanation for the actions made by the system. The reason for the significance of the user interface component is the end-users usually evaluate CDSs based on the quality of the user interface instead of the system itself. The user insert the risk assessment data screening from patient and triangulated with set of rule, such rules are mandatory attributes and optional. The mandatory attribute are (age, BMI, family history, regular physical activity), on secondly important attribute screening remarkable sign of the diabetes such as three cardinal symptoms (excess urine, thirsty and food consumption) the other attribute category as optional. The physician should made decision accordingly the manifestation and risk of patient to inter the system.

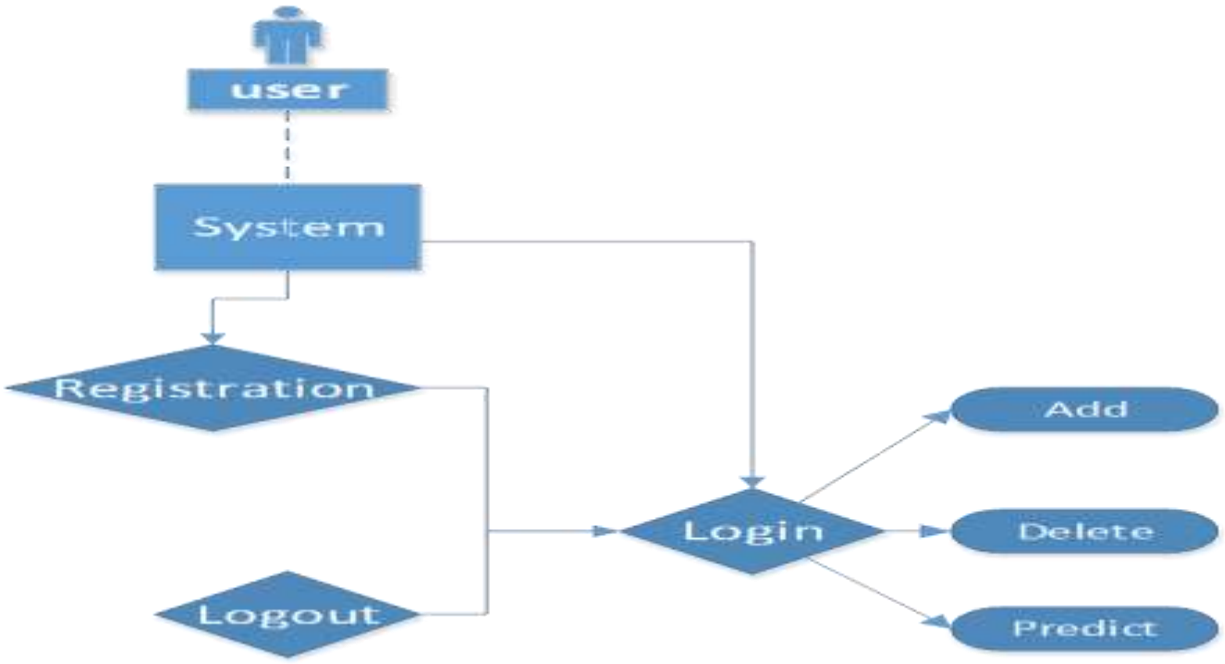

Figure 5.4 User Interface Flow Diagrams.

If the system not included mandatory attribute the interface display error massage, the system satisfy mandatory attribute there is two options either yes or no. yes represent positive for screening displaying massage. And no is negative for screening displaying massage. Home of interface screen display all attributes with drop box with alternative choice, age attribute is take integer and three decision option such as prediction, reset and exit button

Figure 5.5 Graphical User Interface of the Prototype

The second option error massage displayer attribute. This system is displaying error massage the mandatory attribute has not been satisfied.

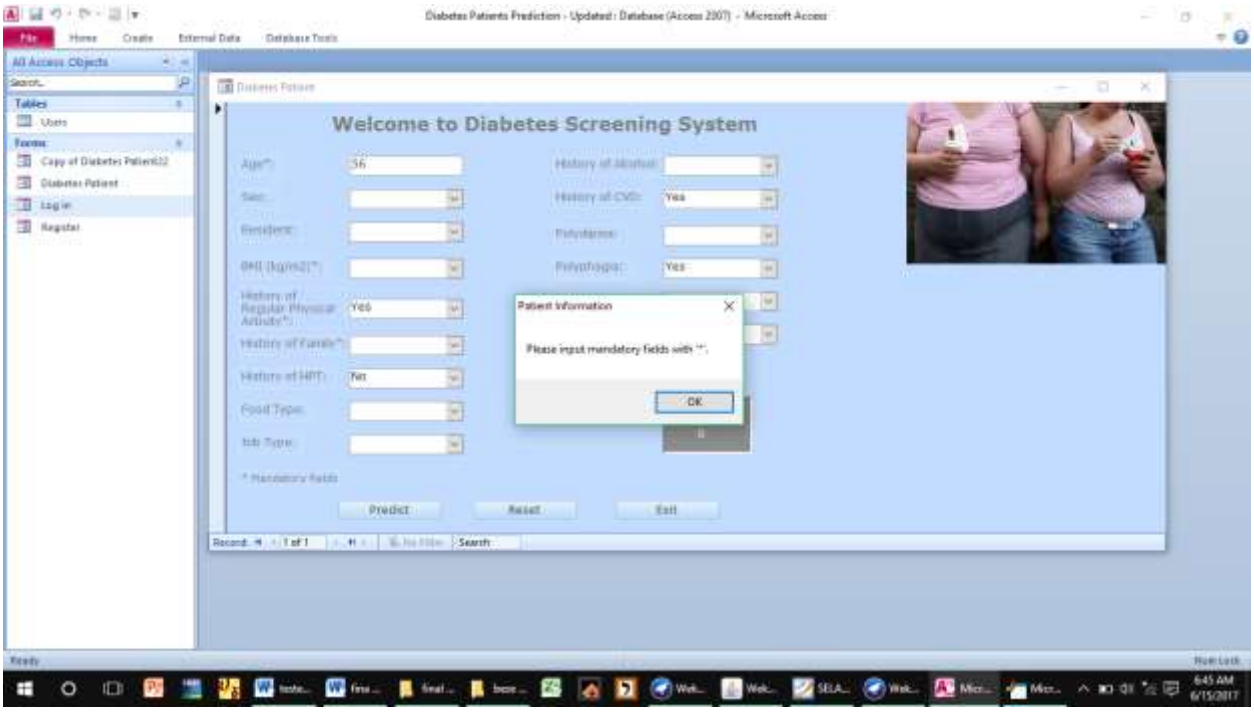

Figure 5.6 Graphical User Interface with error massage unsatisfied mandatory attribute

The other dimension of the interface output displaying interface. The output displaying screen contain results of the prediction either “yes” or positive for risk of diabetes the patient must be referred or link for advanced diagnosis and consultation. The interface display rewarding of choice is screening and alarm to provider discussing the next appointment and counsels the way of the life modification. Or “no” no association risk for diabetes for this patient gives advice periodic checkup and consulted prevention or life modification to dallying getting of diabetes. The result is displaying in following figure.

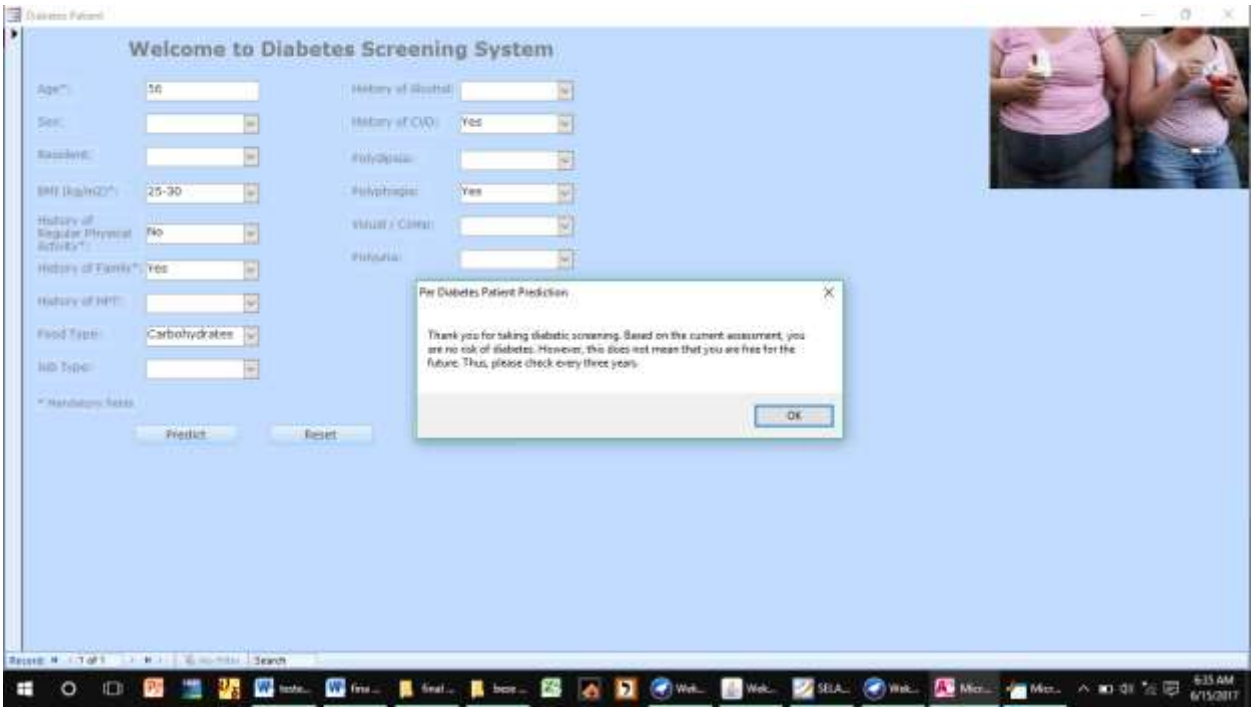

Figure 5.7 display result User Interface with displaying massage

5.5.2 User Interface Testing and Evaluation

After the system was development process complete, the next step was testing and evaluating the system whether the system satisfies the users need and asses’ performance of the system. The scope of testing and evaluation that is accomplished and the significance involved to depend on the complexity, and other core features of the system. As the aim of testing and evaluation of the system is to assure that the system expected what it is required to do. The health professional

selected from diabetes clinic. However, the participants were oriented about the system's flow and what are the system features.

Therefore, in this project testing and evaluation of the system has two aspects. The first one is system performance testing, and the user acceptance testing.

In this study, 10 health professional were selected from Adare general hospital. All health professionals were working in diabetic clinic. The professional tittles of participant are 1 medical doctor, three health officer and 6 nurses qualification and work experience minimum 1 and maximum 9 years to evaluate the usability diabetes screening system. The evaluators of the system were purposively selected. The participants profile includes age, working experience computer skill, and education qualification. Participant information collected sheet attached on the Appendix.

Table 5.5 the participants profile summarized in table

Participant	Age	Qualification	Professional experience	Computer skill
1	28	BSc nurse	4 years	Intermediate
2	25	Diploma nurse	2 years	Intermediate
3	33	Medical doctor	5 years	Intermediate
4	26	Health officer	2years	Beginner
5	30	Health officer	4 years	Intermediate
6	36	Health officer	9 years	Beginner
7	24	BSc nurse	1 year	Intermediate
8	26	Diploma nurse	2 years	Advance
9	30	BSc	6 years	Intermediate
10	27	BSc	3 years	Intermediate

5.5.2.1 System Usability Test

The goal of the system usability test was to determine the usability of diabetes screening system. Usability often refers as the question of how well users can use system functionality. This is also not one-dimensional property of user interface. It's associated with five attributes learnability, efficiency, memorability, errors (error committed rate) and satisfaction. Ten participant tested by

5 question on each participant and the summary presented by no of question time 10 participant divided by 50 mathematically presented as follows.

The result of usability testing is demonstrated as follows. The values for all measurement tools Likert scale in table are fixed as: Strongly agree = 5, Agree = 4, undecided = 3, Disagree = 2 and strongly Disagree = 1.based on evaluation demonstrated five question for the ten selected people 76% responds strongly agree , 16% agree only 8% users are undecided to use the software.

Table 5.6 usability testing of screening model

Usability testing						
Criteria of evaluation	Strongly disagree	Disagree	Undecided	Agree	Strongly Agree	Average
I think that I would like to use this system frequently.				20%	80%	100%
I found the system not complex.			10%	10%	80%	90%
I thought that the system was easy to use					100%	100%
I think that I wouldn't need the support of a technical Person to be able to use this system.			10%	20%	70%	90%
I thought the system doesn't have inconsistency.			20%	30%	50%	80%
Total			8%	16%	76%	**92 %**

Finally, the average usability of the diabetic screening prototype system according to the evaluation results filled by the participants (domain experts) majority of people 92% agreed that the system prototype has a good and clear informational and functional explanation regarding the objective of research project.

5.6. Discussion of Result

As the purpose of this research is screening patient before yet getting diabetes disease. Attributes selection process tried to consult domain knowledge which has strong relationship chronic disease special diabetes, clinician who, working chronic disease follow up department in Hospital to develop predictive model, the findings are discussed in this section. Attribute subset

selection is the process of identifying and removing as much of the irrelevant and redundant information or dimensional reduction as possible. Learning algorithms differ in the amount of emphasis they place on attribute selection.

Such a result are consult with different experts and adoption of WHO screening criteria selected attributes that has highly pertinent to screening diabetes from the 16 features and 1 class (Age, sex, residence, BMI, history of regular physical activities, food type usually eat, family history of diabetes, history of hypertension, alcoholic, job type related to stress, history of cardiovascular disease, attribute related to the symptom of diabetes like often filing hungry, filling thirst, excessive urine and blurred vision related to hypoglycemia and class attribute diabetes status of patient). Finally the evaluation of the selected features with respect to learning algorithms is considered as well it leads to a large number of possible rearrangements(60).

Risk assessment is covering symptoms, recognized life style and behavioral risk factors. In addition to the detection of undiagnosed diabetes there is increasing interest in identifying people without diabetes who are at increased risk of the future development of the condition. Prevalence of diabetes to strongly associated with the risk factor. So attribute selection highly concerned to past exposure of patient. The contribution of developing good algorism screening criteria component on the life style, behavioral factor like smoking, obesity or abdominal obesity, limiting alcohol consumption, preventing high blood pressure. There are also some non-modifiable risk factor age, race or ethnicity, family history or genetic predisposing, the history of gestational diabetes and low birth weight this variables can't be change situation(61).

BMI also risk association is very high, American diabetes association report (2011) female is highest category of body-mass index (over weight) the risk of getting diabetes was more than three times that for men in lowest category (normal weight). Obesity highly associated with DT2, BMI or central obesity correlation with several disease morbidity and mortality. Obesity people are susceptibility increase risk 50–75% developing type2 (NIDDM) diabetes disease, not only obesity but also childhood malnutrition. An increase in BMI is generally associated with a significant increase in incidence of new cases diabetes mellitus. So, based on selected rules and literature survey shows BMI were positively association with developing non-insulin dependent diabetes (NIDD)(62).

The other two classification rules that demand action are identified in terms of age group of patients. Wherever of individual increase age the chance of having diabetes also increase. according to report of American diabetes association (2012) Age group >40 is the most susceptible of the population have had develop T2DM, another study sighted that age is high association risk factor, the risk diabetes in men 55 years old or older was more than twice that in men under 45 years of age(63).

A genetic inheritance of getting diabetes is one of main argument in scientific community. Some studies of identical twins suggest 50% of susceptibility to periodontal disease is due to host factors Genetic predisposing factor; people getting disease from family inherited thought descendants this are contributing factors should also be considered, including genetic susceptibility as a result of variations in, for example, insulin sensitivity(39).

Regular Physical activity decreases the incidence of diabetes nearly by half of. People with a regular physically active lifestyle are less likely to develop insulin resistance of type2 diabetes; the World Health Organization has identified physical inactivity as the fourth leading risk factor for global mortality. WHO report recommended, regular physical activity and weight loss have been shown to delay or prevent the onset of NIDDM and reduce getting of other chronic non-communicable. The set of rules extracted from algorism also significant association with physical activity(63).

5.7 Discussion of the Result on Developed Prototype

The newly developed system does not require additional effort, the project prototype required service integration like other previously existing on health care system such as PICT (professional initiated counseling and testing) integrated with outpatient department and inpatient, STI (sexual transmitted infection) with the adult OPD ,TB screening on all patient with cough for two weeks. In addition to these the new system can provide a screening target group or eligible population mass campaign industry worker, military camp other concentration camp.

The newly developed system is a simple and user friendly. There is no need of advance computer training for managing user interface. The advantage of the screening prototypes save time of patient and health professional investigation and avoid advanced laboratory testing. The

patient risk exposure or clinical information notes of the patient adequate for screening. In addition to this maintain the patient within health care setting introducing periodic checkup and improving health crevice utilization.

The health care professionals evaluated the analysis of developed prototype clinical decision support system, given feedback incorporate with the CDSS. The result showed tested of user interface and performance result 92.5% and 80% respectively. Generally the developed prototype is great contribution for screening pre- diabetes patient then health professional are agreed about the system contributions to the study area specifically and all over the country general.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

High level data computing technology like data mining is improving decision making process for service delivery, store, retrieval and utilization of knowledge, for academic and research purpose. Healthcare sector generate huge data patient records, periodic conducted survey, researches institution data is a good grounds for data mining but, the reason for poor recording and reporting there is a lot of missing the most important elements of data element, results of this hinders decision making process. After all, the structured data's explore hidden value; the potential to predict screening, priority of the program, trends of disease, measure program effectiveness, predict the occurrences of outbreak has largely.

On the other hand, increase in data volume whereas data is form of hard copy and incomplete missing value that cause difficult in decision making process. It is to bridge this gap of inconsistence of data and extracting useful information and knowledge for decision making that the beginning of new generation of data computing methods integrated with decision support system which emerged in recent years.

In order to achieve the objective of the study was conducted according to the CRISP-DM the researcher has been used decision tree and PART rules induction mining. The two commonly used and popular classification algorithms (J48 and PART) and long process of data cleansing, and dimensionality reduction and transformation used it to build the prediction and identify best models. Data mining can be used to help predict future patient behavior and to improve screening and early detection diabetes mellitus prevention program.

Extend objective of this experimental research, which engaged a CRISP methodological approach, made use of two predictive modeling techniques, J48 decision tree and PART, to address the problem. The experiment result shows that PART rules outperformed decision tree classifier. Hence PART with 96.9% accuracy prediction model building was selected to extract interesting rules to develop prototype for achieving the screening program.

The project is achievement of all objectives that application data mining select best model, develop prototype of diabetes screening interface and assessing health care structure and opportunity to implement the project. The result of assessment identify the strategies compatibility with local health care system, no more additional effort unless ones avail computer and train health care professional who working on the adult outpatient department of health facility. Availability of partner closely working on ministry of health and regional health bureau and finally the interface is user friendly easy to operate these and other support project sustainability of the project.

6.2. Recommendation

From the result of the research project findings have implication far elsewhere this. It can be used as one component of a decision support system for diabetic screening in health care organization. The study can contribute May a lot for the further studies conducted in the area of diabetic research. Researcher forwards the following recommendations for future work particularly in relation to diabetes mellitus screening , program support and the effort of the health care system for future.

This study showed the potential applicability of data mining techniques in diabetes screening dataset in developing a classification model. Based on the study, the following recommendations are put forwarded for health managers, Ministry of Health (MOH) and other stakeholders:

- More research and development efforts need to be conducted to enable and explore the variety of data mining techniques that can be applied in diabetes and pre-diabetic dataset.
- Integration of data mining techniques into existing system and computerizing manual recording systems in database is a priority issue.
- Training is highly recommended for data handlers. Therefore, immediate managers of the organization, regional health bureau, MOH and other stakeholders must facilitate conditions for the overall improvement of data handling and storing.
- Besides computerizing the data, consulting experts on recording formats and information to be registered is also a crucial issue in improving the quality of care of patient and health services at all.
- Implementation of the findings primarily in Adare General Hospital and other similar settings.

The size of the dataset has an impact on data mining research. Especially proportional dataset will enhance the performance of the algorithms. Further researches can be conducted using large dataset.

Annex

Reference

1. F, farshad Nba. Cost Effective of type2 diabetes screening medical jorinal of the Islamic Repablic of Iran 2016;30.
2. D, Renuka SM. Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus International Journal of Applied Engineering Research. (2016);11(1):727-30.
3. meeting RoaWHOaIDF. Screening for Type 2 Diabetes. Geneva: 2003.
4. T, Lily MH. Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran. Healthcare Information Resaerch 2013.
5. Srinivasan PRaB, . Predicting Diabetes by cosequencing the various Data Mining Classification Techniques International Journal of Innovative Science, Engineering & Technology. 2014;1(6).
6. B A, E, Akinom On the Diagnosis of Diabetes Mellitus Using ANN Model africa jornal of computing and ICT Referance format 2011;4(2).
7. WHO. Global status report on noncommunicable diseases 2010. 2010.
8. WHO, Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications. diabetes medicince 1998;15.
9. Association AD. Diagnosis and Classification of Diabetes Mellitus. carediabetesjournalsorg 2012;3(1).
10. Federation ID. International Diabetes Federation annual report. 2014.
11. Health FDRoEMo. Health Sector Transformation Plan 2014.
12. Organazetion wh. Non communicable disease country prifile 2014.
13. W, Krishan KHa. Empirical Study on Applications of Data Mining Techniques in Healthcare Journal of Computer Science: . 2006 2((2)):194-200,.
14. D, Tahani AR. Diagnosis of Diabetes by Applying Data Mining Classification Techniques International Journal of Advanced Computer Science and Applications. 2016 7(7).
15. L, Guariguata DRW. Global estimates of diabetes prevalence for 2013 and projections for 2035. diabetes research snd clinical practice 2014;103
16. WHO. Global status report on non-comunicable disease. 2014.
17. WHO. Global Status report on Non-communicable disease. Geneva switzerland WHO, 2010.
18. M.Marco ULa. Limb salvage in diabetic foot: the Italian experience. jornal Diabetic Foot infections: Treatment & Cure 2014.
19. G, Asfawesen GSa. Human Resource Development for Health in Ethiopia Center for National Health Development in Ethiopia. 2007;21 (3):216-31.
20. Haldurai Lingaraj RD, Vidya Gopi , Kaliraj Palanisamy. Predition of Diabetes Mellitus using Data Mining techniques JBC 2015;06(1).
21. Ansari PAN. Prediction and Analysis of Disease through Data Mining Techniques International Journal of Engineering Research in Electronics and Communication Engineering April 2015 2(4).
22. WHO. Deffination , diagnosis and clasification of diabetes mellitus and its complication Geneva 1999.
23. geneva WHO. Prevention of diabetes mellites Report of WHO study group 1994;1(2).
24. J, Han MK. Data Mining Concepts and Techniques. .Stephan A, editor. Morgan Kaufmann Publishers is an imprint of Elsevier 500 Sansome Street, Suite 400, San Francisco, CA 94111: Diane Cerra; 2006 by Elsevier Inc.

25. C, Hian TG. Data Mining Applications in Healthcare. Journal of Healthcare Information Management 2005; 19(2).
26. D, Rajeswara PV, , . Performance Analysis of Classification Algorithms Using Healthcare Dataset International Journal of Computer Science and Information Technologies, . 2015;6(2):1103-6.
27. F, Usama PG. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence. 2006.
28. F, Usama PG. From Data Mining to Knowledge Discovery in Databases. In: magazen A, editor. 1997.
29. Harding ACJ. Data mining in manufacturing: a review based on the kind of knowledge. J Intell Manuf :. 2009;20(501-521).
30. D, Rajeswara PV, . Performance Analysis of Classification Algorithms Using Healthcare Dataset International Journal of Scientific and Research Publications 2016;6(10).
31. Kumar KLaS. Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability. International Journal of Scientific & Engineering Research 2013 4(6).
32. Crockett D. Essential Lessons for Adopting Predictive Analytics in Healthcare. 2013.
33. K.Siri KHa. Empirical Study on Applications of Data Mining Techniques in Healthcare Journal of Computer Science 2006;2(2).
34. G, KHCaT. Data Mining Applications in Healthcare. . Journal of Healthcare Information Management, . 2005. ;19,(2).
35. J, Swati. Cost effectiveness of interventions to prevent and control diabetes: A systematic review 2010.
36. M, Mohammad MMa. Data-Mining Technologies for Diabetes. Journal of Diabetes Science and Technology. 2011; 5(6).
37. Y HaJ, Yingtao A multilayer perceptron-based medical decision support system for heart disease diagnosis. science direct Expert Systems with Applications. 2006;30(2).
38. K.Senthil. A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis International Journal of Advanced Research in Computer and Communication Engineering 2016; 5 (12).
39. S, Sarita JSa. Type II Diabetes a common life style disease Iinternational Jourinal of food and Nutritional Science 2014;3(3).
40. A S, Predicting The Occurrence Of Measles Outbreak In Ethiopia Using Data Mining Technology. 2011.
41. A, Shegaw. Application of data mining technology to predict child mortality patterns: the case of butajira rural health project (BRHP), . 2010.
42. Chhieng JMaC. Data Mining and Clinical Decision Support Systems.
43. S, Vijiyarani SS. Disease Prediction in Data Mining Technique – A Survey International Journal of Computer Applications & Information Technology 2013;2(1):2278-7720.
44. Society IEMaS. International Congress on Environmental Modelling and Software Modelling for Environment's. In: A.Swayne YW, A. Voinov, A. Rizzoli,, editor. Fifth Biennial Meeting, ; Ottawa, Canada 2010.
45. P, Aruna AN. Prediction and Analysis ofDisease through Data Mining Techniques International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE) April 2015;2 (4).
46. Association AD. Diagnosis and Classification of Diabetes Mellitus. america jorinal of diabetic care. January 2012;35(1).

47. R.S, Chandra Sa. Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes International Journal of carrent Research and acadamic rewiev October-2014;2 (10): 91-8.
48. K, Takeshi. New classification and diagnostic criateria of diabetes mellitus by the japan diabetes society. Asian Med J , 2001;44(2):49-56.
49. Organization WH. Diagnosis and clasification of diabetes jeneva suizerland 1999.
50. B**,** Jessica AG, W.Clara , Ayesha A. Motala b. Global estimates of undiagnosed diabetes in adults. Diabetes Research and clinical practice 2014;10(3): .
51. J, sheta Ss. Type2 diabets and common life style disease International Journal of food and nurtition sciences aprl-june 2014;3(3).
52. mistire W, Diabetes mellitus and associated diseases from Ethiopian perspective: Systematic review Ethiop J Health Dev 2013;27(3).
53. B, Nahla APBaBN, . Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus 2011.
54. A.Y, Adekunle. The Prediction, Diagnosis and Treatment of Diabetes Mellitus Using an Intelligent Decision Support System Framework International Journal of Advanced Research in Computer Science and Software Engineering. March 2015; Volume 5(Issue 3).
55. M, Xue-Hui HY-X, R. Dong-Ping Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung Journal of Medical Sciences ,. October 2012;29:93-9.
56. Famili FSW, E.Simoudis, . Data Preprocessing and Intelligent Data Analysis International Journal on Intelligent Data Analysis, . 1997.
57. H, Shafi AM. Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining. Global journal of health science 2015;7 (5).
58. D, Deepti. Comparative Study of Popular Classification Techniques of Data Mining International Journal of Advance Research in Computer Science and Management Studies 2015; 3 (9).
59. M, Ibrahim AM. Reasoning Techniques for Diabetics Expert Systems 2015
60. H, Geoffrey AMa. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transaction on Knowledge and Data Engineering 2003 15,(3,).
61. Harris-hayes Ddam. Epidimology of diabetes and diabetes related complication jornal of the american physical therapy association 2008;88(11):1254-64.
62. Royle MGaP. Non-pharmacological interventions to reduce the risk of diabetes in people with impaired glucose regulation: a systematic review and economic evaluation. Health Technology Assessment 2012;16(33).
63. E Bays aHC. The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. International Journal of Clinical Practice. 2007; 61 (5):737-47.

Annex 2

Survey 1: User Satisfaction

Please circle the numbers which most appropriately reflect your impressions about using this user interface system.

1. I think that I would like to use this system frequently.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	strongly agree

2. I found the system not complex.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	strongly agree

3. I thought that the system was easy to use

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	strongly agree

4. I think that I wouldn't need the support of a technical Person to be able to use this system.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	strongly agree

5. I thought the system doesn't have inconsistency.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	strongly agree

6. I felt very confident using the system

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	strongly agree

System performance test

Please circle the numbers which most appropriately reflect your performance about using this user interface system from never uses to use always.

1. Overall, the system provided adequate functions for screening

	1	2	3	4	5	
Never	○	○	○	○	○	always

2. The system respond to user actions consistent at all times.

	1	2	3	4	5	
Never	○	○	○	○	○	always

3. Are status messages informative and accurate?

	1	2	3	4	5	
Never	○	○	○	○	○	always

4. Overall, the interface was easy to use.

	1	2	3	4	5	
Never	○	○	○	○	○	always

Information Sheet

Good morning/good afternoon.

My name is _____. I came from Addis Ababa University;

I am working for an investigator doing my thesis for the partial fulfillment of master's degree in health informatics. The purpose of this study is designing and developing diabetes screening prototype Adare General Hospital. You will use structured interview template questions to clarify about your risk factor of the diabetes, in order to collect data support designing diabetes screening model.

No personal identifiers will be attached/ recorded to the questionnaire. All the data obtained will be kept strictly confidential by using only code numbers and will be accessed only by the principal investigator. Your participation in the study is up on purely voluntary basis any time you are right withdrawn from the study. What we learn from this study will help to design diabetes screening prototype based on patient collected clinical and epidemiological data. Your honest and genuine participation in responding to the questions prepared is very important and highly appreciated.

If you agree to participate in this study I will give you the information for this study.

Would you be willing to participate? Yes ☐ No ☐

Principal investigator: Bezahegn Zerihun Mob. +251916038059

Email. zbezahegn@yahoo.com

Annex 3 data tool collection designed template

Demographic and clinical characteristics of participants															
Sr no	age	Sex	Mortal status 1=Marred 2=Unmarred 3=Divorce 4=Widowed	Edu status 1= 1st 2= 2nd 3= highr	Residence 1=urban 2= rural	Hx of pyh act Yes No	Family History Yes No	Typ of food Carbo Fat Protein	Hx of HPT Yes N0	Smoking Before Yes No	Job type 1=Gov 2=Non-gov 3=Self 4=Retired	Hx alcohol Yes No	Hx CVD Yes No	Status of pat Diabetes No-diabetes	Type Type1 Type2 No-diabets
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															

Key: - carbohydrate= vegetable, potato, beans, barley and rice

Fat, butter, oil, sun flower, corn animal fata

Protein; all animal products

Data collection tool designed template

Constitutional sign and symptoms of diabetes						
sr.no	Age	Sex	Polyuria Yes=1 No=0	Polydipsia Yes=1 No=1	Polyphagia Yes=1 No=0	Visual complaints Yes=1 No=0
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

DECLARATION

I declare that this project is my original work and has not been presented for degree in any other University and that all sources of materials used for the project have been acknowledged.

Bezehagn Zerihun

This project has been submitted for examination with our approval as university advisors.

Dr. Million Meshesha

Dr. Assefa Seme

Addis Ababa, Ethiopia

June 2017