Prediction of Heart Diseases and Diabetes Using Data Mining Techniques
A Study of MOI Health Center, Kingdom of Bahrain

Ammar Aldallal
Computer Engineering Department, Ahlia University, Bahrain
aaldallal@ahlia.edu.bh

Amina Abdul Aziz Al-Moosa
Information Technology Department, Ahlia University, Bahrain
a.almoosa10@gmail.com

Abstract
Data mining is a relatively new field of research whose major function retrieve knowledge from a large amount of data. It is a process of selecting, exploring, and modeling large amounts of data to discover unknown patterns. Modernization and commercialization of life lead to an unhealthy lifestyle that results in increasing non-communicable diseases like heart diseases and diabetes. Non-communicable diseases have direct result of inaction, inactivity, and idleness of people. Heart diseases and diabetes are two of the most dangerous killers affecting the society. This research aims to produce application software that doctors and other medical practitioners can use to predict the occurrence or recurrence of NCDs. With a huge volume of patient data available, which is not mined for forecasting purposes, a necessary requirement is to utilize this data productively. Data collection and analysis were conducted in two stages: first stage involved secondary data collection and literature review of previous studies. Secondly, primary data was collected to understand the needs and importance of having a tool for doctors to aid prediction of NCDs. Developed software is to be made incorporating relevant secondary data to enable practitioners to use an application which can predict with accuracy, the chances of occurrence or recurrence of NCDs. The research objective is to develop a software application that doctors can rely on for predicting NCDs. The researcher developed a new software application that can use data mining techniques for harvesting patient information and in making predictions.

1. Introduction
The world relies on technology it is incorporated everywhere such as policing, sports, business, engineering, and medicine. Increasing use of advanced computing power with internet has resulted in a global cyber revolution. Information accumulation, analysis, and usage have increased over time. The speed of information sharing and usage has drastically changed how organizations including hospitals functions. Data mining is a relatively new concept used for retrieving information from a large set of data. Mining means using available data and processing it in such a way that it is useful for decision-making.
This project focuses on using huge volumes of data available within the Ministry of Interior (MOI)’s Health Centre; the data is scattered and incoherent. This project focuses on data mining can be used to predict non-communicable diseases (NCD) using data within MOI. According to the statistical figures provided by the World Health Organization (WHO), the International health monitoring organization, the figure shows that globally NCDs are responsible for 68% of fatalities. According to WHO, the major fatal NCDs’ are diabetes, cancer and cardiovascular diseases (WHO Media Centre 2016). In this study, the researcher focuses on Cardiovascular Diseases (CVD) and diabetes. Heart diseases account for twelve million deaths globally, as mentioned in the above source. Especially in developed countries like the US, Europe, Japan, etc., the major contributor to adult mortality is CVDs. CVDs include a whole array of issues relating to the human heart and its functioning. With timely forecasting, by studying patient records and lifestyle, such CVDs can be predicted, and preventable actions can be taken to eliminate / suppress these life-threatening illnesses.

The fashion disease or new generation disease is diabetes. Due to sedentary lifestyle, lack of exercise and bad eating habits like dependence of fast food triggers glucose intolerance. People develop this intolerance to glucose that makes the body produce less insulin in pancreas that is insufficient to control blood sugar. According to WHO, the 2014 figures give a value of nine percent among people aged eighteen and above. One and half million deaths have been attributed to diabetes and related complications.

For an organization like MOI, which staffs more than 23000 employees, the two NCDs Cardio Vascular Diseases and Diabetes present a multi-faceted problem. The problems are not just lost working hours, they highly affect cost of medication, lack of efficiency of staff, and above all contributes negatively to the general health of the organization. Therefore, controlling and reducing the risks associated with these diseases should be the main priority in medical science research. Especially in a law enforcement agency like the 'Ministry of Interior.' Constantly increasing volumes of such critical data can effectively be mined to provide a host of information to decision makers and can aid in predictions for the future. Applying data mining methods to the huge data resource available in the ministry’s database is one of the best ways to utilize information for predicting trends and in identifying susceptible people within the ministry. The problems related to non-communicable diseases are many and costly to diagnose the causes of the illness. These diseases are very serious that if not treated properly and on time can lead to certain complications and even death. Therefore, these diseases are the main priorities in medical science research. Medical diagnosis is an important task that is needed to execute predictions, accurately and efficiently. Unfortunately, doctors are not always experts in every specialization, and there is a shortage of specialized personnel at many places. Hence, any system that would predict or forecast diagnostics (mechanical, semi-automatic of fully automatic) would definitely bring a lot of relief and help to doctors, practitioners, and medical experts (Obenshin, 2004). Software applications and ERPs are widely used by major hospitals globally. Hospitals in Western countries rely on these applications for patient history and health prevention. Unfortunately, MOI does not have a similar software application that can help predict diseases so that they can be treated on time. Hence, Ministry of Interior needs a computerized advanced diagnosis system will prove beneficial. The electronic diagnosis system will gather information for effective prediction of NCD and assist doctors in getting a proper decision-making and support system.

1.1 Purpose of the Study
This research aims to analyze health data such as the vitals to predict NCDs’ using inferential data mining and the use of KDD in order to aid doctors in their analysis of a patient’s potential to have recurring or even predict occurring of NCD. Currently, there are no tools and techniques available within MOI’s Health Centers. Therefore, this research aims to provide a platform for conducting such research in order to develop a ready reckoner type database, available to doctors and other decision makers within the ministry to predict with a certain element of accuracy, the potential for NCDs’ among employees of the MOI. In order to identify the key elements that are needed for predicting NCDs’, an application for compiling the huge volumes of available data on
MOI personnel for NCDs’ for predicting future trends is developed using data mining. Then it will be tested to demonstrate a model whereby doctors and other stakeholders can access and use such information on NCDs’ for improving the health and reduction of such diseases so as to benefit the employees and also MOI. Due to the sensitivity of this project, researchers used certain Healthcare data. The main aim of this research is Creation of a software application using available data within MOI’s database and thereby facilitating easier and accurate predictions of NCD. Secondary data will be utilized, and this is available in the records of MOI. Data will be collected physically, and relevant data will be assembled according to the requirements of the software application. Primary data would be collected after interviewing at least 30 to 40 % of doctors, nurses, and paramedics at the MOI Health Centre.

1.2 Significance of the Study
This project aims to experiment and test different data mining algorithms which will assist in predicting non-communicable diseases and compare and contrast the best possible ways to predict these diseases. This experiment can be a helpful tool for doctors to predict precarious medical cases that involve NCDs and advice their patients according to the predictions provided by the correct algorithm, which will take a very big advantage for the medical science field. This research would prove that data mining technique is a helpful technique for physicians to predict risk cases in their practice, as done elsewhere in the developed world. Section three talks about the type of data used in this research and the analysis of the primary data is aimed at understanding the need for having such an application and also in developing a new application. Section 4 delves into the results of a questionnaire survey done to understand the need and importance of having a computer based application to help doctors to predict NCDs’ and in writing code to develop an application that will be tested. Section 6 contains conclusion and elements suitable for future research.

1.3 Limitations of the Study
This study is limited to the employees of the Ministry of Interior, Kingdom of Bahrain. This research, once completed, would be discussed with Health Authorities in the kingdom and a comprehensive research could then be done to cover the entire population of Bahrain.

1.4 Structure of the Study
The rest of this paper is organized as follows: Section two of this paper is the storyboard where the researcher delves into published information in order to emphasize the aim and objective of using data mining in the medical field. The researcher works for a government health center that caters to 23,000 staff. As of date, there is no available mechanism for doctors or others in analyzing huge volumes of data are lying idle, in terms of patient information. Section three talks about the type of data used in this research and the analysis of the primary data is aimed at understanding the need for having such an application and also in developing a new application. Section 4 delves into the results of a questionnaire survey done to understand the need and importance of having a computer based application to help doctors to predict NCDs’ and in writing code to develop an application that will be tested. Section 6 contains conclusion and elements suitable for future research.

THEORETICAL BACKGROUND AND LITERATURE REVIEW
Data mining is recently gaining international acceptance in almost all spheres of life and medicine in particular. Given the huge scope of data mining in improving health care, this study aims, through this section; to present the theories of determining NCDs’ and the application of data mining in forecasting.

2.1 Overview of Non-Communicable Diseases
One of the most common causes of death in the current modern world is the direct result of the unbridled growth of non-communicable diseases on account of modern day lifestyles and inactivity in people. Physical exercises are mandatory for the maintenance of any system, including the human body. More than 30 million deaths, according to 2005 statistics, were on account of NCDs’ like diabetes and CVDs’. The alarming fact is that more than 50% of this is in people who are less than 70 years old and males and females are affected in a ratio of 1:1 (WHO, 2005).

2.1.1 Diabetes
If we look at statistics recorded by the WHO and compare the figures of 2014 over 1980, a span of 34 years, there is an alarming increase of 409% in the case if living diabetes patients. Standard of living
could be the contributor for such alarming increase because the huge rise in percentage has been attributed to developing or underdeveloped countries (WHO, 2014).

What is diabetes? According to doctors, diabetes occurs when a gland known as pancreas does not release a hormone called insulin in sufficient quantity. Insulin is a hormone that carries sugar from the bloodstream to various cells to be used as energy. Lack of insulin disrupts the body’s natural ability to produce and use insulin accurately. As a result of this, high levels of glucose is released in urine. In the long-term, diabetes when not properly managed can lead to organ failure, cardiovascular diseases and disrupts other functions of the body. WHO has listed diabetes as one of the four major NCDs’ in the world today (World Health Day, 2016).

Statistics released by WHO are alarming. Diabetes, as mentioned earlier, can lead to other major complicated CVDs. According to WHO, 3.7 million fatalities occurs before the age of 70, and this is high mortality rate is attributed to diabetes and CVDs. Uncontrolled blood glucose level is the major factor behind diabetes. With data mining techniques, doctors around the world will be able to predict illnesses effectively and be better equipped to manage potential high-risk candidates. Such analysis and predictions become critical if the objective is to provide relief to millions around the world.

Diabetes is a major health problem in Bahrain. It is the most common chronic diseases across all population and age groups. According to a report by WHO 2016, in Bahrain, 13% of deaths are related to diabetes. This report also shows that the harmful effects of this disease are increasing at a rapid speed as shown in figure 2.1 (Country Profile - Bahrain 2016).

Diabetes is divided into two distinct types; type 1 diabetes enforces the need for artificially infusing insulin through medicines or by injections and type 2 diabetes, pancreas create insulin, but it is not effectively used by the body. The majority of people with diabetes are affected by type 2 diabetes. Diabetes was a common problem among adult’s specifically middle-aged people but due to changing lifestyles diabetes effects children too.

Type 1 diabetes is unpreventable because of the various external environmental stimulants which result in the destruction of body’s insulin-producing cells. However, changing lifestyle to achieve the required body weight and obtain the physical activities can help to prevent type 2 diabetes to enlarge.

According to International diabetes federation, diabetes could be prevented by:

1. It is recommended to begin a weight loss regime to curb the effects of diabetes by improving insulin resistance and reducing hypertension. Therefore, people with overweight should encourage, achieve and maintain a healthy body weight.
2. Continuous physical activity is essential to maintain the weight loss. Physical activity reduces hypertension, arrhythmia, and sensitivity to insulin, bring an improvement in body compositions and improve psychological health.
4. Smoking, stress or depression, and sleeping patterns may also consider as risk behaviors, which preferred to be improved to prevent diabetes.

Although, physicians need to know how to quickly identify and diagnose potential cases. By strengthening the ability to access, collate, analyze
2.1.2 Cardiovascular diseases

Cardiovascular diseases are a group of disorders of heart and blood vessels such as:

- Coronary heart disease: clogged, narrowed or blocked blood vessels that supply blood to heart muscles cause this heart disorder.
- Cerebrovascular disease: problems in arteries that can affect the blood supply to the brain causes cerebrovascular disease. Stroke is an example of cerebrovascular disease.
- Peripheral arterial disease: caused by plaque (fats, calcium, and other substances) built up in arteries that supply blood to limbs, head, and other organs.
- Rheumatic heart disease: streptococcal bacteria cause this heart disease, which affects connective tissues in body specifically in heart, and brain.
- Congenital heart disease: is caused by structural anomalies and malformations present at birth.
- Deep vein thrombosis and pulmonary embolism: this disease is caused by blood clotting in leg veins which can be transferred to other organs like heart and lungs.
- Myocardial Infarction (MI) is the medical term for what is one of the most common killers NCDs’ which is known better as Heart Attack. Cerebrovascular Accident (CVA), also called stroke is yet another serious NCD. These illnesses occur due to the malfunction of heart and its arteries. The arteries are vessels that pump purified blood into the body from the heart. On account of bad eating habits, lack of exercise, accumulation of fat in the body leads to these arteries getting deposited by fat cells in the inner walls of the arteries. Use of tobacco, misuse of alcohol, hypertension, irregular eating habits and a host of situations lead up to MI or CVA.

The buildup of fat inside arteries and blood vessels lead to clogging of the vessels in the course of time and as time progresses, without medical care or treatment or changes in eating habits or changes in lifestyles, can ultimately lead to fat cells completely clogging the artery or vessel leading to a complete blockage of flow of blood. Unmanaged blood glucose levels, obesity, and physical inactivity or leading a sedentary lifestyle are common in today’s population. Previously, people used to walk a lot, work in the fields and farms, fish in the sea and involve a lot of physical labor which in today’s world is dwindling fast. Man, in his quest for material gains and comfort, has invented millions of things that reduces the need for relying on physical activity. The advent of the fast-food culture is yet another huge danger that many of us fail to understand or turn a blind eye to. Once a person is detected to have hypertension, cholesterol imbalance, diabetes, etc., it is time to ensure that such person gets informed on the impending dangers that would lead to NCDs’. Such “intermediate risk factors” as described by the World Health Organization can be identified by even a basic health center or clinic.

It is common knowledge to the literate world that stopping or reducing the dependency on alcohol or tobacco, limiting greasy foods that are high in fats and reducing sugar and salt intake can aid people to reduce the risks associated with NCDs’ and CVDs’ in particular. When a potential threat is detected, people must immediately seek medical help and reschedule their lifestyles. Moreover, people should use appropriate medication to control or limit the damage caused by these factors.

In most of the cases, there will never be a clearly visible symptom for a stroke or an MI. It happens, all of a sudden and people are caught unaware. In fact, most MIs happen when people are asleep. When people experience pain or discomfort in the chest or shoulders, or excruciating pain that shoots up in the right arm or CVA. Immediate medical attention is the only option in this scenario.

Most common symptoms associated with CVAs’ are numbness or insensitivity of arms, legs or face and it could even be one entire side of a body. There could be disorientation of speech possibly even difficulty in vision. Severe headaches or dizziness or hallucinations can happen, and in extreme cases, patients might even faint.

The world, which is populated by low and middle incomes groups, is the most affected by NCDs. The WHO records speak for themselves. The rich have access to medical care but the low and middle-income groups, especially if we look at country profiles, have little or even no access to basic medical care. These are countries across Asia, Africa and Southern America where there are no health centers within reach. As a direct for such groups, which
account for the majority of human population, such ailments are detected very late so that treatment and rehabilitation cannot work. NCDs can further complicate issues for the low/middle-income groups. The father of the house, the only bread winner for the houses if suffering from an NCD, cannot go out to work or earn and this leads to further increase in poverty and mal-nutrition.

According to WHO fact sheet (2014), cardiovascular diseases caused 26% of deaths in Bahrain. The statistics indicate that the probability that 13% deaths within the age range 30-70 years will be caused by four NCDs. Physicians must monitor risk factor of cardiovascular diseases to prevent reasons that cause heart attacks or strokes by getting the right system for storing the medical record and analyze the data.

2.1.3 Diabetes and Cardiovascular Disease: Double Jeopardy
Diabetes is a major contributor for CVDs, as proven by clinical trials and cases of people suffering from CVDs. Due to malfunction of critical organs like the kidneys and liver, it becomes more strenuous for the heart to pump blood. Malfunction of the liver, kidneys, pancreas, and dangerous amounts of toxins unflushed thus remain in the blood stream. Medical world has to a challenge to deal with this double-edged sword. This is the correct time for Country’s medical centers and the governments to look closely at how to contain the two killers: diabetes & CVDs. To help to prevent deaths by CVDs data mining techniques should be applied to forecast with reasonable degrees of accuracy the occurrence or reoccurrence of diabetes & CVDs.

2.2 Data Mining, Meaning, Theory, and Evolution
Over the last century, the use of advanced technology has helped develop tools and machines that would help in faster and accurate computing more efficiently and more accurately. From the basic abacus, man has invented computational tools from calculators to modern day, microprocessors that can collate, analyze and present data in nanoseconds. If we look at organizations now, every day they generate a huge volume of data which could be specific to the organization in order to rationalize and use such huge volume of data man invented algorithms in order to micro-focus data to super specific requirements. This further led to creation and development of machine learning algorithm that would help in generating different types of analysis and decisions with minimal human supervision. Data mining thus has evolved based on human needs which can help humans in identifying relationship patterns and forecasts based on pre-set rules and stipulations built into the program (Eapen, 2004).

Different authors and practitioners use different terms to describe data mining which will be explained further in this section. From data mining evolved the concept of discovering knowledge from databases. Frawley and Piatetsky (1996) describes data mining as the process of extracting implicit and previously undisclosed important information about data sets that can be used for effective decision-making. The process is termed as Knowledge Discovery in Database.

Such discovered knowledge can be very useful in many areas of sciences, and health care is no different having a KDD would help in predicting trends of many kinds of diseases and illness. So doctors, rather than depending on their own knowledge and experience, can use data mining and specifically KDD to predict or to forecast and to predict trends that would lead to better diagnoses, reduce cost and save person-hours for the organization. Another key aspect of KDD is increasing efficiency and effectiveness of doctors who can constantly treat large numbers of patients at a given point of time. Similarly doctors of the same profession who could properly be in different organizations, locations or even countries. Doctors can share and consult online to provide the best diagnoses possible within the least amount of time. At first, we should understand the growth and factors affecting heart attacks and diabetes before building predicative models. Then understand the data mining techniques and algorithms.

The utility of data mining can be seen from different angles. Otherwise, seemingly irrelevant data can be logically aligned to bring out hidden correlations and patterns. Such correlations and patterns can provide invaluable information that can prove critical to a person’s health and well-being. The idea is to summarize huge volume of data into useful information (Clifton, 2010, Friedman, Jerome (2009).
Data mining is placed as a statistical interface. Data mining lies in the interface of statistics, database technology, recognizing patterns, machine readable data, and intelligent expert systems (Obenshain, 2004). The prime objective of data mining is to extract information from data sources and alter it into a comprehensible assembly of information for more uses (Data Mining Curriculum, 2014). Data mining is a process that is used to locate correlations between data and form pattern of relationships among cluster fields in the enormous interactive database (Extract-Nature Biotechnology 18, 2000).

According to Krishnaiah, Narasimha and Chandra (2016), data mining is a major facility that enables the use of data and databases by techniques used for recognizing and utilizing patterns or trends for the data set. According to them, the key aspect of data mining is to find patterns or trends mechanically or automatically, with least user input and efforts.

Today the study of data mining is modern enough to encompass concepts of artificial intelligence, effectively performing fast-paced database management technology and visualization (Yoo et al., 2011). As cited in Kamber’s (2010) book “Data Mining: Concepts and Techniques” define data mining as the method that can be used to extract hidden information in data with ease, the data being true and accurate which has not been disclosed previously and comprises of important information in massive data depositories.

The author contends that data mining is known by other names such as knowledge discovery in databases (KDD). Using KDD as a tool for knowledge gathering, data analysis, and data archeology, and harvesting large amounts of essential information. According to Usama, Gregory and Smith (1996) argues that more than raw data analysis, data mining databases can be used for modeling, visualization, and online updating. So according to others, data mining is one step of the KDD which we can call the collection and analysis stage.

Jiawei and Micheline (2001) argues that the term data mining is an incorrect word as the objective is to define previously unknown trends from available databases. This according to them is not actually mining of data, rather utilizing available data in a scientific way to extract trends and patterns. As stated in a conference at Arizona State University (OKAIRP, 2005), data mining is the process of the huge volume of data and applying decision making systems which can lead to logical solutions. According to Witten, Frank, and Mark (2011), the term data mining was originally named ‘practical machine learning.’ However, the introduction of the term data mining is considered as a marketing tactic.

What is the database? As the term implies, the databases and accumulation of historic data which can be made available and can be updated for an organization or individual use (Obenshain, 2004). Most databases have a provision for asking questions that can elicit logical information, and if we look at today’s world, the scope of databases is increasing exponentially. According to Obenshain (2004) who published a work on data mining application to healthcare data contends that data mining methodology may be adopted on huge volumes of data and using a computer. In the past, people relied on statistician who sat and computed volumes of data manually. Data mining incorporates using new variables apart from its ability to use large quantities of variables.

The real data mining tasks are related to the mechanical analysis of massive amounts of data which is used to acquire information that was previously not disclosed. Data mining helps in pattern identification and categorizing data records by conducting cluster analysis, identification of odd records also called detecting anomalies and association rule mining or dependencies. One of the database techniques used in the spatial index (Peter, Witten, Ian, 2010).

Various statisticians’ communities of information systems and data analysts have used the term data mining. The KDD is known as the complete procedure used to acquire data or knowledge discovery. Data mining is a core concept in KDD. Therefore, data mining can be termed as an application that uses precise algorithms that are used to identify and extract data patterns. Data mining is considered as the focal center of KDD. It deals with applications of intellectual approaches that are applied for extraction of data patterns. From a healthcare point of view, traditional methods as statistical process control based on numerous fundamental probability distribution functions are successfully implemented and evaluated to control infections in hospitals (Benneyan, 2001).
There was a time, as early as the 1960s when mathematicians and statisticians looked down upon data mining. They used the terms ‘data fishing’ or ‘data dredging’ because they considered it a bad practice in the world of mathematics, the idea of analyzing data without a set hypothesis. In the 1980s, the term database mining was coined by a San Diego-based company, HNC to describe their Database Mining Workstation (Jesus, 2011). By 1990s, researchers seem to have accepted the term data mining as a legitimate phrase to describe the method of harvesting available data to predict future happenings. In fact, data mining was called by many other names: Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, etc. (Jaishankar, 2015). Gregory Piatetsky Shapiro coined the term Knowledge Discovery in Databases or KDD. However, the term data mining gained relevance and acceptance in areas of artificial intelligence and machine learning communities, and further, this term was popularized in the business and the fourth estate communities (Shapiro, Gary 2011). Currently, the terms data mining and knowledge discovery are interchangeable after 2011, data mining was described as data science. According to the academic research community, a major forum for research in data mining was conducted in 1995. This research was carried out during the first Data mining and knowledge discovery conference (KDD-95) in Montreal. The co-chair of the conference Usama Fayyed, launched the first journal in 1996 and it was called the Journal of Data Mining and Knowledge Discovery and then a SIGKDD Newsletter (Usama 1999).

2.3 The Knowledge Discovery in Databases (KDD) Process:
Knowledge discovery in databases (KDD) is the procedure used to attain important and useful knowledge from a large collection of previously collected data. The process involves selecting, preparing and cleansing the data from unnecessary information. Any previously available information is incorporated into the data sets. Data interpretations are conducted to achieve precise outcomes from available results.

As cited by Oded Maimon and Lior Rokach from “Introduction to Knowledge Discovery.” The KDD procedure initiates with determining goals and objectives, and the process ends by implementing the new knowledge that was discovered during the process. The loop ends there, and the data mining segment initiates. Resulting in alterations in the application domain (such as offering varied features to cell phone users for reduction in churning). The loop ends with this process, and all outcomes are measured over the new data sources. After all, steps are completed the KDD process restarts. Below is a brief explanation of the KDD process, as elaborated by Kamber (2010):

Step 1. Having a clear idea of the application domain
Understanding what the end-user, the doctors or medical experts need is of primary importance and the clarity of though in what is “REALLY” expected should be the starting point of the KDD process. Here the data miner gets a clear-cut vision on what is expected and then define what needs to be done to reach the desired objectives. There can be changes in the eventual goals or objectives as the end-user might want additional answers to related questions but this is normal. Once the required knowledge is decided, the people involved in the KDD should now do the preprocessing exercise which is briefly explained in the following three steps.

Step 2. Selection
This stage necessitates the definition of the goals and the details of the types and quantities of data that are to be used that will be further processed to achieve the desired goals. This is again, a very critical stage
because the faulty selection of data sets can lead to further complications on the way and can lead to irrelevant or useless inferences in the end. The end user's job would be further complicated by getting wrong information or wrong inferences which can lead to wrong inferences by the doctors, which in turn can harm a potentially helpable patient. Any model that is to be created relies heavily on the selection of the datasets and its collation. In order to be on the safe side, the authors recommend the selection of more than really necessary attributes for the excess of data would not lead to a deficiency in the output.

Step 3. Preprocessing and cleansing.
Before the identified set of data is further processed, the authors recommend that the said data is subjected to a filtering and cleansing stage. The idea is to improve and augment the reliability of the chosen set of data. Noise or barriers that may impede this data needs to be removed at this stage. These stages may use a lot of time and effort, but at the end, one can be assured of the validity, reliability, and utility of the identified data. It can happen that sometimes, the certain data set can be insufficient or inadequate. In such situations, where miss leading and / irrelevant data can interfere with the correct prediction, care must be taken to ensure that such gaps are filled and that the algorithm is sufficient and enough to be processed for the expected and desired output. For example, irrelevant patient data can act as a barrier so in this stage such barrier should be removed.

Step 4. Data transformation.
Making the data to be project specific is the next step in that to ensure that the best set of data is available so as to ensure the best possible outcomes or results which can aid the doctors to make logical and meaningful predictions and diagnoses. This is a very crucial step in any KDD project. As an example, say, in sales, when predicting future sales, one need to make a provision to accommodate events that are unnatural or beyond control, otherwise, sudden shocks can shake up the organization's targets. Likewise, in sports, one need to make provisions for uncontrollable that may adversely or positively influence or affect a game. If provisions are not made at this stage, events can happen that can even fully destabilize or scuttle the KD process. Dimension reduction and attribute transformation are two of the methods suggested by the authors. The KDD process thus is self is a platform for self-introspection and possible future transformations that may be needed to ensure the validity and efficacy of the outcomes.

Step 5. Choosing the appropriate Data Mining task
Now comes the job of choosing the type of data mining to be selected for the particular type of project. This choice depends on the end result expected or the goals of the project and also to the stages that were explained above. Choosing between regression and classifications or clustering depends on the desired output. The normal outcomes of a data mining exercise can be two: 1. Forecasting and 2 the outlay or description. Forecasting or prediction means data mining is done under controls of supervision and the other talks about data mining done under visualization.

Step 6. Selecting the right Data Mining algorithm
According to the authors, this stage now defines the tactics to be employed to achieving the strategic objectives. Will using neural networks suffice? Alternatively, is using decision the best option? The choice depends on the type of searching pattern that fits ideally to the given project and the desired end-result. Is it precision of the result or the decipherability or understandability of the result that the project needs? Neural networks are better according to the authors when the precision of results are needed, and the decision tree is appropriate when understanding of patterns and trends is required.

Step 7. Employing the Data Mining algorithm
This is the step of data mining algorithm by employing the algorithm many times until the best result is obtained. For example, by adjusting the algorithm’s parameter controller such as the minimum number of occurrences in a particular leaf, accuracy of the desired result is obtained.

Step 8. Evaluation
Evaluation of the mined data is the essence of this stage in the KDD process. Evaluation means making reasonable interpretations and findings with respect to the defined and pre-set project goals and objectives. We may need to add or delete features in the data transformation stage. The usefulness and
ability to be comprehended by the end-users is achieved at this stage. Once finalized, the newly found knowledge is documented. Once the KDD team finalizes and documents the whole process of data mining, the stage is now for testing the application.

Step 9. Using the discovered knowledge
The effectiveness and efficiency of the developed KDD process will now be tested by the end-user and refinement can still be done if fine tuning is desired. In effect, the created knowledge is now being pilot tested by doctors and practitioners to ensure that the desired outcomes are being generated and that it helps in better diagnoses and medical forecasts when applied to MOI’s health center. When we test a new product in actual environments, there can be situations that might differ from laboratory conditions. There must be built-in flexibility to adapt, change, or modify the new knowledge so that the actual end user stands to gain from the new knowledge.

2.4 Data Mining Techniques
Nowadays data mining attracts a great deal of interest in two industries. These industries can be divided into two categories:
Corporate activities and E-commerce
Scientific, Engineering and health care
There are two types of Data Mining tasks the predictive model and descriptive model. Both models are explained as follows:

2.4.1 Predictive Model
The predictive data-mining model predicts the future outcomes based on past records present in the database or with known answers. This model is used by many organizations that try to data mine a person’s credit worthiness. For instance using a car loan application system for determining if a client is a negative or positive credit risk. Data mining will help figure out the future credit risk of the applicant and predict future credit history of the applicant by using past data.
Classification is known as the procedure used to locate a model that best suits identified data sets or ideas. The model helps predict the class of objects when class labels are not available. The resultant model is focused on analyzing a set of identified classes.
Regression is a mathematical and statistical tool used widely in using numeric values for forecasting.
Time series analysis
Prediction as the term implies means correctly envisioning the future using logical computation of available data.

2.4.2 Descriptive model
This model is to discover patterns in the data and understand the relationships between the data attributes. Descriptive Model represents the main feature of the data, and summarize it. For example, marketers can discover diverse sets in their customer databases. The collected knowledge can be used to develop marketing programs for targeting audience. Clustering examines data objects without referring to an identified class label. Summarization is to categorize the distinctive properties of data and point out if the data values are to be categorized as noise or outliers. Association rules, the associations, and patterns among several characteristics are obtained by the same rules recreated. These patterns and rules are employed for predicting the classifications of data which was being tested.
Sequence Discovery This is the systematic order identification of events or occurrences that are registered either in a timely fashion or randomly. The researcher classifies data mining as shown in the figure below.

Figure 2.3 - Data mining tasks
Out of these models, we will be only concentrating on one method, which is a prediction-mining method.

2.5 Data Mining in Medical Applications, Forecasting

2.5.1 Healthcare Informatics

The definition of Medical Informatics is dynamic due to the fast paced and changing nature of both medicine science and computing technology. As cited by Hoyt and Yoshihashi (2014) healthcare informatics is a discipline that deals with materials, resources, tools and formal ways used for storage optimization, managing and retrieving biomedical information needed for decision making and problem-solving. Similarly, Lazakidou and Siassiakos, (2009) highlights that healthcare informatics is using computer applications, modern communication and IT and computer systems implementation in medicine fields like medical care, education, and research. Moreover, according to Bronzino (2006) health care informatics is applying skills by using different tools that help using and sharing information for better health care delivery and promote health. Another term for health care informatics is Medical Informatics. While some practitioners term is Clinical Informatics or Bioinformatics. However, Bioinformatics includes integration of technology and biology. Then Bioinformatics is defined as the computer and statistical based analysis of biological information that is to be used in databases and algorithms to speed up and improve biological research. Clinical Informatics is responsible for maintaining structuring, organizing data, searching data and making decisions using relevant data will be useful in conducting important medical research.

2.5.2 Data Mining in Medical Field

Over the last few decades, the tremendous gains and inventions that have happened in the world of computing and the computer related analyses and inferences have benefitted mankind in general, and the field of medicine is no exception. Enormous gains have been made to clinical diagnoses on account of software and high-performance computing. Previously, doctors and medical practitioners relied on their own knowledge, skills and above all intuition. Today, for example, cardiologists from different organizations, different regions, and even different countries have formed formal or informal links so that they can share clinical information and statistics which can be used by the rest of the world for the benefit of mankind. Data mining is but one of the major gains that the medical field has benefitted from when it comes to forecasting NCDs’ and major diseases, their occurrences and recurrences. The fact is that data mining now has become an integral part of any forward-looking medical agency, and it is not by choice but a norm that health care facilities around the world are being encouraged to utilize the tremendous potential and use of data mining in forecasting disease trends and occurrences. Information management in healthcare is an area where data mining has proven to be of huge importance and value. According to Acharya and Wu (2010) the numerous algorithms developed and used in data mining has tremendously aided healthcare agencies, health-related informatics, patient monitoring systems, epidemiology among a whole array of utilities. The differentiation of pathological data from normal data has significantly aided medical practitioners in diagnoses and decision making. According to Craus and Allori (2009) and Srinivas (2010), the creation of knowledge in medical systems can hugely be aided by using the technique of mining available data and in making coherent inferences from that. Cause and effect relationships between medical disorders, symptoms, and effects can be compared, collated and analyzed to make predictions by creating appropriate software. Especially in predicting cardiovascular diseases, data mining has proven to be tremendously useful and successful. The study aims to find frequently used information items using candidate generation. The process will involve taking massive databases as input for the software application by effectively utilizing the expansibility of the software application. By using proper tools, information will be retrieved from these databases by exploring hidden knowledge. The dataset related to most commonly observed diseases in medical databases can help reduce the datasets of these illnesses by data mining process. It is essential to have all information related to risk factors accompanying heart diseases, and diabetes can assist doctors in effectively diagnosing patients with high risk of getting NCDs.

Modern data mining techniques along with essential statistical tools can help doctors to diagnose diabetes...
and cardiovascular diseases. By using statistical analysis, several cardiovascular diseases such as cardiac arrests (coronary heart disease), stroke (cerebrovascular disease), hypertension, congenital heart diseases, arrhythmia, peripheral artery disorder and heart failure with diabetes. Cardiovascular diseases are caused by unhealthy lifestyle and eating habits, chewing or using any kind of tobacco, fast food ingestion, and alcohol abuse is the main reasons behind these illnesses (Santa Kumar et al., 2010).

Today, elsewhere in the world, a technique called the artificial neural network technique is used as a mining method to effectively predict and forecast cardiac arrest situations. After pre-processing and clustering the available dataset before use (Ambarasi, 2010). Subsequently, the identified significant trend is collated with the neural network training. Multi-layer Perceptron Neural Network with Backpropagation is used for training. This proved beyond that the said algorithm is really effective in forecasting heart attacks.

The performance results of decision trees are more accurate and have lesser drawbacks as Bayesian classification. Other classification methods have not been found to give such accurate results. Heart disorders can be predicted by using 15 attributes along with basic data mining approaches like Artificial Neural Network (ANN), Clustering and Association Rules, soft computing approaches, etc. K-Nearest Neighbor, Neural Networks, Classification based on clustering will not be a good performer (Khaing, 2011). Another study conducted by Acharya and Wu (2010) who analyzed the dataset of 900 odd CVD patients and tested the attributes for consistency found that after applying genetic algorithm found that the decision tree analysis worked out better and more efficiently with almost 100 percent accuracy.

In 2011, HninWintKhaing presented a well-organized approach for predicting cardiovascular diseases and resulting cardiac arrest risk level. In the first phase cluster of all cardiovascular diseases is formed by using K-Means Clustering algorithm. The algorithm will help acquire relevant data related to cardiac arrest from the database. This method allows understanding the number of fragments through k-parameter. Consequently, the recurrent patterns are mined from the extracted data that are relevant to heart problems by using the MAFIA (Maximal Frequent Item set Algorithm) algorithm.

The machine language algorithm is acquainted with using selective yet important patterns for efficiently predicting cardiac arrests. By using ID3 algorithm as the main training algorithm to represent the designated level of cardiac arrest and disease prediction system. The system is effective is predicting cardiac arrests by decision trees (Chaurasia and Pal, 2013). A research study was conducted to effectively predict cardiac arrest risk levels by using the cardiac arrest risk level database. Forecasting the occurrence and recurrence of CVDs necessitates the utilization of 11 attributes with KDD techniques like Naïve Bayes, J48 decision tree, and Bagging approaches. The results indicate that using bagging approach is much more accurate as compared with Bayesian classification and the J48 (Quinlan, 1986).

Researchers are constantly applying different algorithms and methods like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, and Nearest Neighbor methods to help doctors diagnose a heart disease or diabetes effectively so that further damage can be prevented. The above extract and of lots of other researchers prove beyond doubt, the growing influence and power that data mining has in the future of prediction and correction of NCDs’.

2.6 Summary
Almost all models developed using data mining techniques are theoretical models in many countries they have adopted these theoretical models to create a useful application for their country. There is no such application in Bahrain, and the researcher aims to create a practical model that is specific to Bahrain and its unique demographics.

The Literature review provides an overview of the study in terms of previous studies conducted on the same topic. The reader is introduced to the theories of data mining, non-communicable diseases and the use of technology like data mining to elicit knowledgeable and realistic predictions on NCDs’. Internationally, data mining has been in use by the medical field for decades. At the Ministry of Interior’s Health Center, where the researcher works, does not have any application or tool for realistic and timely forecasting of common forms of non-
communicable diseases, especially diabetes and cardiovascular diseases.

3. RESEARCH METHODOLOGY
The major aim of mining prediction approach is to design an effective predictive model that can classify the most effective factors that require good input data (parameter), with suitable data mining model. In the research data mining will be used to create, class based selected attributes prediction models. The research has three main factors mainly patient database, prediction methods for data mining and implementing an application which will be used to categorize and develop a prediction model for diagnosing cardiac arrest and diabetes risk in patients included in the dataset taken from MOI’s patient database. This research uses both primary and secondary data for the study. The goal of this study is to utilize available knowledge on data mining, its importance in healthcare and its uses in the medical field and then to develop an application software which would be tested in real life.

3.1 Types of Data
3.1.1 Secondary data
Secondary data, which means already published information has been collected, read and used from articles, journals, books and other published data like hospital records. The secondary data is incorporated into the literature review and further used in the development of the application software.

3.1.2 Primary data
The researcher would develop a complete set of codes using preset parameters to design a program. The information used in this software comprises the information from the Ministry of Interior’s Health Centre database which is presented in Section 4, in the patient data set. The purpose of this study is to develop an application software that will help medical experts to predict the chance of occurrence of non-communicable diseases among staff of the Ministry of Interior, Kingdom of Bahrain. The four aims of this section are to explain the research methodology, explain data collection methods and techniques, describe the application software development and to explain different tools and procedures used for data analysis.

This study uses secondary data, in the beginning, to describe the concept of data mining and how it is used in medical inferences. Numerous articles and publications were used to collect and write up the literature review, which explains the need and importance of using data mining techniques to the benefit of forecasting the chances of occurrences of non-communicable diseases.

3.2 Procedure for Data Collection
Data for primary research is compiled by administering a questionnaire survey and also by developing a software program using available secondary data in the health center records.

3.2.1 Primary Data Collection
A questionnaire survey among doctors to understand clearly about their opinion and ideas about the current situation regarding NCDs and the utility of having a software application to help them predict NCDs’. To investigate the current scenarios regarding prediction of NCDs’, the Ministry of Interior’s doctors who are either endocrinologists or cardiologists, a questionnaire survey will be conducted. The survey is one of the most commonly used data collection tools to gather relevant data from a sample selected from the target population. Usually, a questionnaire is used in surveys that the sample participant needs to complete. The interview is another effective tool for data collection. (Robson 1993).

Surveys are conducted to gather data from people about their perceptions and opinions about a certain topic. The survey is the most preferred data collection tool for information analysis from chosen sample respondents. They can be effective in both quantitative and qualitative data collection methods in social sciences (Rossi, Wright, Anderson, 1983).

3.2.2 Population and Sampling
As on date, there are 13 cardiologists and 17 endocrinologists. The entire population would be the sample as the number is relatively easy to work around. Hence the sample would be all the concerned doctors. The researcher chose a personally administered, structured questionnaire to find out the views and opinions of serving doctors who are either endocrinologists or cardiologists at the Ministry of
Interior’s Health Centre that is responsible for around 23,000 employees.

3.2.3 Data Analysis
In order to truly understand the views of the practitioners on the efficacy of having software that would help them better their diagnosis, a questionnaire survey was conducted. The aim of the survey was to find out the importance which, doctors give to such an application and the utility of having one such. Questionnaire survey results are to be tabulated and percentages or ratios to be computed to find out the awareness / utility of predicting NCDs’ among all endocrinologists and cardiologists within MOI Health Center. Program coding is to be done by the researcher and pilot-tested before being applied.

3.3 The Proposed Application Software
The researcher applied a technique of data mining for predicting heart and diabetes risks for individual patients of the Ministry of Interior. The researcher used the mining comparison prediction algorithm and used patient data sets attributes that affect the prediction. The program was developed using VB.Net coding language and Oracle as the database.

3.4 Summary
This section discusses the research methodology, and it gives a detailed description of the types and sources of data (both primary and secondary) and the modules of using these data to develop an application software which would aid doctors in the MOI’s Health Centre in better predicting occurrence or recurrence of non-communicable diseases.

4 ANALYSIS AND FINDINGS
The researcher, having gone through the process of collection, collation, and analysis, describes the analysis and the findings. This section explains the model of the questionnaire used during the survey. The survey ensured that the ultimate end user, the doctors who will be using the application are convinced on the importance of having such an application in predicting trends in NCDs’ within MOI. Model for developing the software is explained in detail, which is the desired output of this research. The research emanated from a necessary requirement within MOI’s Health Center for a way of diagnosing NCDs’ like diabetes and cardiovascular diseases are the most common forms found in patients.

4.1 Questionnaire Analysis
Even though the researcher was certain that such a software application is not available in MOI or in Bahrain, to validate this assumption, a questionnaire survey was conducted on the sample. This is done by doctors to ascertain the current situation as far as usage of tools in forecasting NCDs. The total population of Endocrinologists and Cardiologists who are currently on the rolls of the Ministry of Interior is the full sample to be tested. Since interviews were conducted on a one-to-one basis, there were no missed interviews. All 30 doctors who were approached completed the questionnaire.

After compiling the data provided by 30 doctors; 13 Cardiac specialists and 17 endocrinologists (many are specialized but working as General Practitioners) the results tabulated and the analysis is as follows:

On the query on approximating the number of Cardiovascular diseases and diabetes patients account for 70% of the total number of patients these doctors treat. This proves that NCDs’ and more specifically cardiovascular and diabetes patients are the majority. Common illnesses like cold, influenza, allergies, sprain etc. account for 30%. Health Center does not treat critical illnesses like cancer.

Question 2 relates to the major causes for such NCDs’ and the doctor's rate bad eating habits, smoking, lack of exercise and eating high-fat food are the major reasons attributed for both NCDs.’

3&4. These questions are related and hence analyzed together. Although 80% of the doctors are aware of the existence of some form of NCDs’ prediction mechanism in the world that they are aware of, they categorically say (100 %) that there is no such mechanism or facility available in the Ministry of Interior’s Health Centre that caters to around 23,000 employees.

5&6. 24 of the 30 doctors polled (which is 80%) of the population agree that it would definitely help if there is a computer-based application which can forecast/ predict such NCDs’ and such information would aid the ministry in alerting patients on a continuous basis.

60 % of the doctors polled (18/30) have previously used similar applications, elsewhere.
90% of the doctors polled (27/30) welcomes the idea of having such an application that can “correctly predict” the occurrence of NCDs and that it is essential.

4.3 Application Description – New Software for Predicting NCDs

An application developed in this research to show the use of patient data values to predict its health risk. This will help in understanding the use of data mining predicting techniques. The database used is explained in section 3 and it is related to the heart diseases and diabetes of patients' data. The database was obtained from Ministry of Interior Health Center.

4.3.1 Designed Application

The objective of this application is to predict whether the patient is at risk of getting a non-communicable disease or not from data obtained from Ministry of Interior Patient Data.

The Tools that have been used to program the application as follows:

VS.Net: The program that has been used to write the application
VB.Net: Is the coding language that has been used to write with.
Oracle DB Is the Database language that the Ministry of Interior have
Toad Oracle Is the tool that used to collect the data

After reading about data algorithms the researcher conducted the right algorithm to be used in this application which is mining class comparison, this concept means that to mine a description that compares or distinguish one class from another comparable class.

The procedure of mining comparison, in general, is as follow (Han, Kamber, 201):
Data Collection. It is the set of useful and relevant data that is gathered using query processing. The data is segregated in a target class with a set of contracted classes.
Dimension relevance analysis. The dimension relevance analysis is to be applied if there are many dimensions, to select only highly relevant dimensions for further analysis.
Synchronous generalization. Generalizations are to be applied to the target class to the extent of control by the user. For outcomes for a prime target class relation. The concept of contrasting classes is to generalize the same level as those in the prime target relation, forming the prime contrasting classes’ relation.

Presentation of the derived comparison. The resulting class comparison description can be visualized in the form of tables, graph, and rules.

4.3.2 Patient Data Set

With the help of the proposed data set, the patterns help to predict the risk of getting heart diseases and diabetes that were extracted using the discussed data mining algorithm.

All patient data sets were already processed by eliminating duplicate patient records and entering the missing values. Furthermore, the most recurring patterns are mined using the recommended data mining process. As listed in table 4.1 below, these are the attributes that have been collected from the data and its cut-off values and description.

<table>
<thead>
<tr>
<th>No.</th>
<th>Attributes</th>
<th>Description</th>
<th>Cut-off Value</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Age of the patients in years</td>
<td>Age &lt;=40, Age &lt;=60 and &gt;40, Age &gt;60</td>
<td>Young Age, Middle Age, Old Age</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
<td>Gender</td>
<td>1, 0</td>
<td>Male, Female</td>
</tr>
<tr>
<td>3</td>
<td>CP</td>
<td>Chest Pain Type</td>
<td>Value1, Value2, Value3, Value4</td>
<td>Stable, Angina, No-Stable, Angina, Non-Angina Pain, Asymptomatic</td>
</tr>
<tr>
<td>4</td>
<td>Test bps</td>
<td>Resting blood pressure (in mmHg)</td>
<td>BP&lt;80, BP&lt;90, BP&gt;90</td>
<td>Normal, Normal to High, High</td>
</tr>
<tr>
<td>5</td>
<td>Chol</td>
<td>Serum Cholesterol (mmg/dl)</td>
<td>Chol&lt;5, Chol&lt;5.2, Chol&lt;5.2 and</td>
<td>Normal, High, Severe</td>
</tr>
</tbody>
</table>
After the preprocessing, the attributes mentioned previously in table 4.1, are used in the application code to use the IF condition. The attributes classified as the cut-off values to do a comparison.

4.3.3 Description of the application process

The entire exercise was conducted hand in hand with the doctors who will be using this software. This is to ensure that there is no irrelevant or unnecessary data that may act as barriers to predictions of NCDs'. The patient data set have been selected from three tables by doing the second step of the KDD process, as shown in figure 4.1, first table is to get the Patient information record, the second table is for the vitals of the patients, blood pressure, diet, smoking, chest pain type and Heart Rate. The last table is the blood test collected from the lab which the attributes are the cholesterol and fasting blood sugar. The tables would have been better as a compiled one but the volume of information in the database would make such a compilation, messy.

<table>
<thead>
<tr>
<th>No.</th>
<th>Attributes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>FBS</td>
<td>Fasting blood Sugar</td>
</tr>
<tr>
<td>7</td>
<td>Rest ECG</td>
<td>Resting electro cardio graphic results</td>
</tr>
<tr>
<td>8</td>
<td>Diet</td>
<td>On healthy diet</td>
</tr>
<tr>
<td>9</td>
<td>Smoking</td>
<td>Smoker patient</td>
</tr>
<tr>
<td>10</td>
<td>Alcohol</td>
<td>Alcohol drinker</td>
</tr>
</tbody>
</table>

Table 4.1 Attributes and Values

Figure 4.1 - ER Diagram of the application

Considering that the vitals and the lab results are the target class and the if condition rule as the constructing class, and the attributes are the Age, Sex, CP, Test bps, Chol, FBS, Rest ECG, Diet, Smoking, Alcohol as shown in table 4.1. The data mining task can be shown in Data Mining Query Language DMQL as follow:

Use Patient_DB
Mine comparison as “Patient_NCD_Risk_Level”
In relevant to Age, Sex, CP, Test bps, Chol, FBS, Rest ECG, Diet, Smoking, Alcohol
For “Vitals_View” and “Lab_Results’ Versus “if_condition_rule”
Analyze sum
Display as “risk level”

First, the query is changed into two interrelated queries that accumulate two sets of relevant task data: one for the target class from the Vital_View table and Lab_Result Table, and the other is from the contrasting class which the “if condition rule” that have been written as classes in the code. Second, the dimension relevant analysis can be performed on the two tables, Vitals_View and Lab_Results. And the weakly relevant dimension such as patient information is removed from the resulting test classes. Figure 4.2 shows all values required to predict the risk level.
Third, Synchronous generalization is performed on the target class to the levels of the controlling class generating the prime target class relation. By using the If condition, we classify attributes in conditions then compare them as shown in figure 4.3. Firstly, compares the ages, if the age is less than 40 it will group it as a young age, if the age between 40 and 60 it will be group it as a medium age, and if the age is greater than or equal 60 it will group it as an older Age.

As done in the below figure 4.5, BP grouped as group 0 if it is less than 80, Grouped as 1 if its greater than or equal 80 and less than or equal 90, and grouped as 2 if its greater than 90.

The below figure 4.4 shows the comparison of patients if smoker or drinks alcohol. If the patient is a smoker goes to group 1, if not group 2. If the patient drinks alcohol goes to group 1, if not goes to goup 2.
Figure 4.6 - Grouping Heart Rate

In figure 4.7, the cholesterol grouped as 0 if the value is less than 5.2. or grouped as 1 if the value is greater than or equal 5.2 and less than 6.2. And if the value is greater than or equal 6.2 then it grouped in 2.

Figure 4.7 - Grouping Cholesterol

Figure 4.8 contains the list of chest pain condition. According to the program if the chest pain is grouped as 1. It will relate that the patient is suffering from chest pains while if the chest pain is grouped as 2 that will signify a risk for cardiac arrest. If the chest pain signal is 3 then patient might be suffering from coronary artery disease. If the chest pain grouped as 4 it means that there is a potential high risk of chest pain.

Figure 4.8 - Grouping Chest Pain

Finally, the resulting class comparison is presented in the form of rules as shown in table 4.2. For example, there are no risks for the patient if all classes are equal to 0.

4.4 Test Results, Predictive Performance of the Application

As described recently the result conditions is the comparison between prime target classes and the applied if condition rule as shown in the below table 4.2, where the result shows as a rule of the risk results.

<table>
<thead>
<tr>
<th>BP</th>
<th>Heart Rate</th>
<th>Cholesterol</th>
<th>Chest Pain</th>
<th>Age</th>
<th>Risk Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>No Risk</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>No Risk</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>No Risk</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Low Risk</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>No Risk</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>No Risk</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>Low Risk</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>Medium Risk</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Low Risk</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>Medium Risk</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Low Risk</td>
</tr>
</tbody>
</table>
Table 4. 2 - Comparison of Classes and Risk results

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2 or 3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2 or 3</td>
<td>2</td>
</tr>
</tbody>
</table>

The screenshots (Figure 4.9 – 4.12) prove the efficacy and utility of the program in predicting NCDs’ for the ministry’s employees. Theses samples of NCDs prediction are shown for different risk levels (Low, Medium, High).

Figure 4.9 - Sample of NCD Prediction Level: No Risk

Figure 4.10 - Sample of NCD Prediction Level: Low Risk

Figure 4.11 - Sample of NCD Prediction Level: Medium Risk

Figure 4.12 - Sample of NCD Prediction Level: High Risk
4.5 Findings

Since there were two analyses for primary data, the researcher wishes to present the inferences under separate headings:

Inferences from the analysis of the questionnaire survey conducted with doctors:

The questionnaire survey proved that currently there are no mechanisms or tools available within the Ministry of Interior’s Health Centre which caters to 23,000 employees to predict or forecast NCDs’. Huge volumes of data are available in the form of patient records, but there is no analysis of such important data for predictions. The survey points out that most of the doctors are familiar with such tools being used elsewhere, and they would want to have such an application available which, would aid them better in forecasting a patient’s chances of recurrence or occurrence of NCDs’.

Inferences from the analysis of the developed software:

The software was tested with 10 patients, and the results are given as screenshots based on predicted risk levels: low risk, no risk, medium risk and high risk. The findings were discussed with doctors, and they are extremely happy to have this tool as told by them when the researcher asked them to test it out themselves. Since NCDs’ can lead to potential mortality among patients, the doctors who tested the software feels strongly for having such an application as it would aid them tremendously in their diagnosis. In fact, they would want to use them immediately for they strongly feel that it aids quicker and better forecasting while briefing patients or in informing potentially high-risk people with a track record.

As envisaged from the very beginning, the development of a software application to predict NCDs’ among men and women who are employees of the Ministry of Interior was the expected outcome. From literature reviewed and based on interviews conducted with practicing doctors, the need was felt and amplified for the development of such software. The creation of software which uses theories of data mining has enabled the researcher to prove that such software can be beneficial for forecasting NCDs’ based on available medical history would definitely aid the ministry and also the staff in better monitoring NCDs’ and in cost savings and time savings for the ministry. To have a fully efficient and effective workforce, especially in a critical organization like the Ministry of Interior, which is responsible for the security and concord of this country, Bahrain, there must be an honest effort to utilize available data and to mine it to predict human diseases with a certain amount of accuracy. The research has proven the utility of having a computer-based software application that can be at the fingertips of doctors or all other concerned people, which can be used, to predict occurrences or recurrences of two of the most common NCDs’ that are found in the modern world.

CONCLUSIONS AND FUTURE WORK

An Application using a data mining algorithm of classes’ comparison has been developed to predict the occurrence of or recurrence of NCDs’ like heart diseases and diabetes risks. In addition, the result of the application shows that the predictions system is capable of predicting NCDs’ diseases effectively, efficiently and most importantly, timely. That means the application is capable of helping a physician in making decisions towards patient health risks. It generates results that make it closer to the real life situations. That makes the data mining more helpful in the health sector, which means that it is necessary for knowledge discovery in the healthcare’s sector.

The study proves that Ministry of Interior’s Health Center which caters to more than 23,000 employees must have a system or a software application that can predict NCDs. The system of application should have a large amount of otherwise unutilized patient / employee medical data that could be mined in order to make logical, timely and precise predictions. The prediction will be useful not just for doctors, but also for the benefit of employees to maintain a check on their health and well-being. The program that has been developed should be tested on a continuous basis, and the coding should be perfected so that the application can be adopted across the Health Center and all clinics under the health center.

Much more than huge savings in costs in terms of medical expenses, loss of duty time and usage of critical medical facilities, this application can provide a one-stop-shop for all MOI employees to have a kind of read-reckoned to have firsthand information about their health and well-being. This application would be a tremendous asset for doctors who can
have structured, specific and invaluable information about their patients / others so that they can ensure that their diagnosis or inferences are correct and professional. In addition, such information can be used internationally. For example, if a patient travels abroad and has a health issue, the person’s complete medical history and his NCDs’ predictions can be made available online, or at the very least, can be mailed electronically, in a few minutes.

There are different algorithms that can produce more applications to help to predict data. A possibility for future work could be to implement an application for continues prediction or a dynamic prediction for more numbers of patient’s health. Furthermore, future work could be considered to do as a mobile application where the patient can enter their vitals in the mobile, with the rule set the result appears right away. This will be a handy tool for healthcare industries.

Finally, the huge appreciations received from the doctors on having such software prove that in a country like the Kingdom of Bahrain, where NCDs’ are on the rise, such applications should be developed to cover the entire country. The common person stands to benefit from doctors having such a tool so that he/she can be better knowledgeable as far as personal health and wellbeing is concerned.

REFERENCES


[12] Han, Jiawei., and Kamber, Micheline., (2001), Data Mining: Concepts and Techniques, pp. 5


[16] Lazakidou A. Athina and Siassiakos M. Konstantinos, Handbook of Research on Distributed Medical Informatics and E-health, 2009


[19] Obenshain, Mary, K., Application of Data Mining Techniques to Healthcare Data,
Statistics for Hospital Epidemiology, (August 2004)


[28] Yoo, Illhoi, Alafaireet, Patricia, Marinov, Miroslav, Pena-Hernandez, Keila, Gopidi, Rejitha, Chang, Jia-Fu and Hua, Lei, Data Mining in Healthcare and Biomedicine: A Survey of the Literature, Med Syst DOI, Springer 2011