

The Islamic University-Gaza
Deanship of Research and Graduate Studies
Faculty of Information Technology
Master of Information Technology



الجامعة الإسلامية - غزة
عمادة البحث العلمي والدراسات العليا
كلية تكنولوجيا المعلومات
ماجستير تكنولوجيا المعلومات

Prediction of Myocardial Infarction by Data Mining among Chronic Diseases Patients (UNRWA Clinics in Gaza Strip as A Case Study)

توقع السكتة القلبية بين المرضى ذوي الأمراض المزمنة بواسطة
التنقيب عن البيانات
(عيادات وكالة الغوث الدولية في قطاع غزة كدراسة حالة)

By
Ahmad H. Elkahlout

Supervised by
Dr. Ashraf Maghari

A thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Information Technology

June /2020

إقرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Prediction of Myocardial Infarction by Data Mining among Chronic Diseases Patients (UNRWA Clinics in Gaza Strip as A Case Study)

توقع السكتة القلبية بين المرضى ذوي الأمراض المزمنة بواسطة التنقيب عن البيانات
(عيادات وكالة الغوث الدولية في قطاع غزة كدراسة حالة)

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الآخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

I understand the nature of plagiarism, and I am aware of the University's policy on this. The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted by others elsewhere for any other degree or qualification.

Student's name:	احمد حسن محمود الكلوت	اسم الطالب:
Signature:		التوقيع:
Date:	7th June 2020	التاريخ:

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة عمادة البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ احمد حسن محمود الكحلوت لنيل درجة الماجستير في كلية تكنولوجيا المعلومات/ برنامج تكنولوجيا المعلومات وموضوعها:

توقع السكتة القلبية بين المرضى ذوي الأمراض المزمنة بواسطة التنقيب عن البيانات (عيادات وكالة الغوث الدولية في قطاع غزة كدراسة حالة)

Prediction of Myocardial Infarction by Data Mining among Chronic Diseases Patients (UNRWA Clinics in Gaza Strip as A Case Study)

وبعد المناقشة التي تمت اليوم السبت 13 ذو القعدة 1441 هـ الموافق 2020/07/04م الساعة الحادية عشرة صباحاً، في قاعة اجتماعات كلية تكنولوجيا المعلومات اجتمعت لجنة الحكم على الأطروحة والمكونة من:

.....
.....
.....

مشرفاً ورئيساً

مناقشاً داخلياً

مناقشاً خارجياً

د. أشرف يونس مغاري

د. توفيق سليمان برهوم

د. سامي سليم أبو ناصر

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في كلية تكنولوجيا المعلومات/برنامج تكنولوجيا المعلومات.

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله تعالى ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق،،،

عميد البحث العلمي والدراسات العليا



د. بسام هاشم السقا

التاريخ: 5 / 8 / 2020م

الرقم العام للنسخة

237571

اللغة

ع

ماجستير

دكتوراه

الموضوع/ استلام النسخة الإلكترونية لرسالة علمية



قامت إدارة المكتبات بالجامعة الإسلامية باستلام النسخة الإلكترونية من رسالة

للتالبة/ محمد محمود الكحلوت

رقم جامعي: 120153049

قسم: تكنولوجيا المعلومات

الكلية: لسانها للدراسات

وتم الاطلاع عليها، ومطابقتها بالنسخة الورقية للرسالة نفسها، ضمن المحددات المبينة أدناه:

- تم إجراء جميع التعديلات التي طلبتها لجنة المناقشة.
- تم توقيع المشرف/المشرفين على النسخة الورقية لاعتمادها كنسخة معدلة ونهائية.
- تم وضع ختم "عمادة الدراسات العليا" على النسخة الورقية لاعتماد توقيع المشرف/المشرفين.
- وجود جميع فصول الرسالة مجمعة في ملف (WORD) وآخر (PDF).
- وجود فهرس الرسالة، والملخصين باللغتين العربية والإنجليزية بملفات منفصلة (PDF + WORD).
- تطابق النص في كل صفحة ورقية مع النص في كل صفحة تقابلها في الصفحات الإلكترونية.
- تطابق التنسيق في جميع الصفحات (نوع وحجم الخط) بين النسخة الورقية والإلكترونية.
- ملاحظة: ستقوم إدارة المكتبات بنشر هذه الرسالة كاملة بصفحة (PDF) على موقع المكتبة الإلكتروني.

351

والله والتوفيق

توقيع الطالب

إدارة المكتبة المركزية

محمد محمود الكحلوت

Dedication

This work is dedicated to the soul of my father and is also dedicated to my dear mother, my thanks for her because of being there when I always needed her. It is also dedicated to my brothers and sisters for their prayers and support. Special thanks to my better half, my wife for her patience, encouragement, and endless support.

Acknowledgements

This year is the year that I have completed my master thesis having the educational and life experience, first of all Alhamdulillah for giving me the strength to accomplish this mission.

I'm grateful to my supervisor Dr. Ashraf Y. Maghari, without his help, guidance, and continuous follow-up; this research would never have been.

I would like to thank the academic staff of the Faculty of Information Technology who helped me during my master's study and taught me different courses.

Without forgetting my deepest thanks to the officials in UNRWA Clinics in Gaza Strip for full collaboration.

Finally, I am greatly indebted to my family for their support during my course studies and during my thesis work.

Abstract

Background: Patients with non-communicable chronic diseases especially cardiovascular diseases have a high rate of morbidity and mortality worldwide. Still, Myocardial Infarction (MI) is the major cause of death and disability universally, in spite of the advanced cardiovascular medical services in the last two decades. It is believed that an early prediction of myocardial infarction among cardiovascular patients is a crucial goal for international health systems. **Objective:** This study aimed to assess the predictive ability of MI using data mining techniques among patients with chronic diseases. **Methods:** Data for this study were collected from five primary health care centers of UNRWA in Gaza Strip. Data for about 8001 chronic patients were processed by using three data mining algorithms (decision tree (DT), K-Nearest Neighbor (KNN), NAIV Bayes (NB)). The algorithms processed the data to extract the probability of a heart attack from combined risk factors available for each patient regarding his health status. Data was reduced by reduction technique for obtaining a reduced representation of the data set. Relevant and attributed features were included e.g. age, gender, smoking status, obesity, history of hypertension, diabetes, lipid profile, and family history of heart attacks. On the other hand, associated complications such as previous MI and brain stroke were excluded from the framework of data mining for more precise sensitivity and accuracy of prediction. Data were transformed into appropriate forms for mining such as data discretization. **Results:** The highest accuracy of prediction was recorded by the KNN algorithm with 97% followed with NB (93%), while the DT reported 92%. Data mining algorithms are able to predict MI among susceptible patients. **Conclusion:** The accuracy of the DT, KNN, and NB techniques were relatively similar, however, the highest accuracy was for KNN. Medical concepts should be considered in selection of patients' features before processing data.

ملخص الدراسة

الخلفية: يسجل المرضى الذين يعانون من أمراض مزمنة غير معدية وخاصة أمراض القلب والأوعية الدموية نسبة عالية من الوفيات في جميع أنحاء العالم. ومع ذلك، فإن السكتة القلبية هي السبب الرئيسي للوفاة والعجز على مستوى العالم، ذلك على الرغم من الخدمات الطبية المتقدمة للقلب والأوعية الدموية في العقدين الماضيين. يعتقد أن التنبؤ المبكر بالسكتة القلبية بين مرضى القلب والأوعية الدموية هو هدف سامي للأنظمة الصحية الدولية.

الهدف: تهدف هذه الدراسة إلى تقييم القدرة التنبؤية بالسكتة القلبية باستخدام تقنيات استخراج البيانات بين المرضى الذين يعانون من أمراض مزمنة. تم جمع البيانات لهذه الدراسة من خمسة مراكز للرعاية الصحية الأولية للأونروا في قطاع غزة. **المنهجية:** تم معالجة بيانات حوالي ٨٠٠١ مريضًا من ذوي الأمراض المزمنة باستخدام ثلاث خوارزميات استخراج البيانات (Decision Tree، KNN، NAIV Bayes) تم معالجة البيانات بواسطة الخوارزميات لاستخراج احتمالية الإصابة بنوبة قلبية من عوامل الخطر المجتمعة المتاحة لدى كل مريض حسب حالته الصحية. تم تخفيض البيانات عن طريق تقنية التخفيض للحصول على تمثيل مخفض لمجموعة البيانات. تم تضمين الميزات ذات الصلة مثل العمر، والجنس، وحالة التدخين، والسمنة، والتاريخ المرضي لارتفاع ضغط الدم والسكري، ومعدل الدهون في الدم، والتاريخ العائلي للنوبات القلبية. في المقابل، تم استبعاد المضاعفات المصاحبة مثل سكتة قلبية سابقة أو جلطة دماغية من إطار استخراج البيانات للحصول على مصداقية أعلى للقدرة التنبؤية. تم تحويل البيانات إلى أشكال مناسبة مثل عملية تقدير البيانات. **النتيجة:** سجلت خوارزمية KNN أعلى دقة للتنبؤ بنسبة ٩٧٪ تليها Naïve Bayes 93. خوارزميات استخراج البيانات قادرة على تنبؤ الجلطات القلبية بين المرضى المعرضين. **الخلاصة:** كانت دقة شجرة القرار، KNN، وتقنيات NAIV Bayes متشابهة نسبيًا. ومع ذلك، كانت أعلى دقة لـ KNN. يجب أن تؤخذ المفاهيم الطبية في الاعتبار عند اختيار ميزات المرضى قبل معالجة البيانات.

TABLE OF CONTENT

DECLARATION.....	I
DEDICATION.....	II
ACKNOWLEDGEMENTS.....	III
ABSTRACT.....	IV
ملخص الدراسة.....	V
TABLE OF CONTENT.....	VI
LIST OF TABLES.....	IX
LIST OF FIGURES.....	X
LIST OF ABBREVIATIONS.....	XI
CHAPTER 1 INTRODUCTION.....	1
CHAPTER 1 INTRODUCTION.....	1
Background.....	1
Myocardial infarction.....	2
Problem statement.....	3
Objectives.....	3
1.1.1 Main objective.....	3
1.1.2 Specific objectives.....	3
The importance of the study.....	3
Scope of the study.....	4
Research Framework.....	4
Thesis structure.....	5
CHAPTER 2 THEORETICAL FOUNDATION.....	7
2.1 Data Mining.....	7
2.1.1 Introduction.....	7
2.1.2 The KDD Process.....	8
2.1.3 Data mining functionalities.....	10
Classification.....	11
2.2 Classification Algorithms.....	13
2.2.1 Naive Bayes.....	13
2.2.2 k-Nearest Neighbors.....	15
2.2.3 Decision Trees.....	16
2.3 Myocardial Infarction.....	18

2.3.1 Major Types of MI.....	19
2.3.2 Symptoms of MI	20
2.3.3 MI diagnosis	20
2.4 Summary	21
CHAPTER 3 LITERATURE REVIEW	22
CHAPTER 3 LITERATURE REVIEW	23
3.1 Introduction.....	23
3.2 Cardio Vascular Risk Factors	23
3.2.1 Age.....	23
3.2.2 Sex	24
3.2.3 Smoking	24
3.2.4 Family history of cardiovascular disease.....	25
3.2.5 Lipids	25
3.2.6 Triglycerides	25
3.2.7 Low density lipoprotein	26
3.2.8 High density lipoprotein	26
3.2.9 Obesity	26
3.2.10 Physical inactivity.....	26
3.2.11 Hypertension.....	27
3.2.12 Diabetes	27
3.3 MI and BMI with traditional statistical method:	28
3.4 Diabetes and Data Mining:	28
3.5 MI with data mining.....	29
3.6 Summary	33
CHAPTER 4 METHODS AND EXPERIMENTS.....	34
4.1 Collect and preparing the data set.....	36
4.1.1 Data set attributes.....	36
4.2 Experiments Setup:	38
4.2.1 Experimental Environment:	38
4.2.2 Experimental Tools:.....	38
4.3 Data preprocessing.....	39
4.4 Data mining methods	42
4.4.1 Decision Tree	43
4.4.2 KNN	46

4.4.3 Naïve Bayes	49
4.5 Summary	52
CHAPTER 5 RESULTS AND DISCUSSION.....	53
5.1 Decision Tree:	54
5.2 K-Nearest Neighbor	59
5.3 Naïve Bayes:	59
5.4 Results conclusion	61
5.5 Discussion	62
5.6 Limitations of the study	63
5.7 Summary	63
CHAPTER 6 CONCLUSION AND RECOMMENDATIONS	64
6.1 Thesis summary	65
6.2 Conclusion	65
6.3 Recommendations	66
6.4 Future Works	67
REFERENCES	68

LIST OF TABLES

Table 3-1 Comparison between the main article related to this work	31
Table 4-1 :Chronic diseases patients instant count	36
Table 4-2:Chronic diseases patient's data set attributes	37
Table 4-3:Excluded Factors	38
Table 5-1:Decison Tree Rules Table	56
Table 5-2: KNN Accuracy with different K Value	59
Table 5-3: Distribution for MI in fields with more satisfactory density.	60
Table 5-4: Accuracy List	61

LIST OF FIGURES

Figure 1.1: Research Framework.....	5
Figure 2.1 The world is full of data and poor knowledge	7
Figure 2.2 : Data mining - searching for knowledge in data	8
Figure 2.3 :The Process of Knowledge Discovery in Database.....	9
Figure 2-4 : The Data Classification Process	12
Figure 2-5 : Decision tree the mammal classification problem	16
Figure 2-6 : History of Documents on the Definition of Myocardial Infarction..	19
Figure 5-1: Myocardial Infarction (MI) decision tree output result.....	55
Figure 5-2: Part of Decision Tree predictor.....	58
Figure 5-3: Decision Tree accuracy.....	58
Figure 5-4: KNN Accuracy with K value 4.....	59
Figure 5-5: naive bayes accuracy.....	60
Figure 5-6: Graphical representation of the accuracy, Precision and Recall for each classifier.....	61

LIST OF ABBREVIATIONS

ACS	Acute Coronary Syndrome
BMI	Body Mass Index
CAD	Coronary Artery Disease
CRP	C-reactive protein
CVD	Cardio-Vascular Disease
DM	Diabetes Mellitus
DM	Data Mining
DT	Decision Tree
ECG	Electrocardiogram
HTN	Hypertension
KDD	Knowledge Discovery Data Mining
KNN	K-Nearest Neighbors
LD	Lipids Disorders
MI	Myocardial Infarction
NB	Naïve Bayes
Patient DOB	Patient Date of Birth
SVMs	Support Vector Machines
UNRWA	The United Nations Relief and Works Agency
WHO	World Health Organization

CHAPTER 1
INTRODUCTION

CHAPTER 1

INTRODUCTION

Background

In the last decades, using IT has led society to generate large amounts of data in all of our life aspects such as business, marketing, science, economics, and medicine etc. Usually, there is knowledge hidden in the data which when extracted can be used for decision making. In healthcare environment, “Data can be a great asset to healthcare organizations, but they have to be first transformed into information(Chen, Chiang, & Storey, 2012). There is a huge amount of clinical and administrative data available within healthcare systems. Knowledge Discovery in database (KDD) refers to the “Non trivial extraction of implicit previously unknown and potentially useful information about data”, the core of KDD is the data mining, which is used for exploring large datasets to extract hidden and previously unknown patterns, relationships and the difficult to detect knowledge with traditional statistical methods (Yeh, Cheng, & Chen, 2011).

Chronic diseases or non-communicable disease defined as Known as his condition, which usually lasts for 3 months, may last more, and if it continues, it gets worse over time. This type of disease, which is known as chronic disease, is considered to be more likely among the elderly, and it will control and coexist with it, but cannot cure it. There are many chronic diseases, and what is known is that cancer is one of them. Also, can say heart diseases, also diabetes, stroke and arthritis. It has a large proportion of human diseases. Of the total deaths expected in 2016, 30.6% were from cardiovascular disease as the first cause of death, followed by cancer which comes in the second place with 14% (PAIC, 2017).

In areas with low or limited incomes, it has a bigger problem than others. Of great importance is the rapid and early intervention of this epidemic or to predict occurrence may be prevented there is many factor as The high blood pressure and the lack of properly organized fat, smoking and the interference of psychological factors and social attachment and high sugar also and many factors represent myocardial infarction in both sexes and also includes all ages does not apply to adults without

children. A preventive approach that is uniform throughout the world has the potential to prevent MI cases before they occur.

this study aimed to assess the ability of data mining techniques to predict MI among non-communicable patients.

Myocardial infarction

Myocardial infarction (MI) considered one of the most common causes which leads to death and morbidity and including a high cost of care (WHO, 2002). There is no doubt that the prediction of this type of disease, known as myocardial infarction, is useful in preventing its development and that the development of artificial intelligence has helped to identify the voluntary history of many diseases, including diseases of the heart.

The term MI leads to one of the major cause of death and disability all over the world (Denmark et al., 2019). There is a lot of research at the end of the last century that showed that there is a relationship between the emergence of (MI) disease and a blockage in the coronary artery, the first clinical descriptions appear in the beginning of the 20th century as that there are clinical features associated with it related to the appearance of a clot in the coronary artery. Regardless of the observations, it has been a long time before the clinical acceptance of this year was achieved, due to a study conducted during the autopsy and she said that there are no blood clots in the arteries, especially coronary ones, who suffer from (MI) during the deceased patients. By time, several MI definitions have been used, which leads to controversy and confusion. Till the WHO He defined myocardial infarction disease and this definition depends on the planning of the electrocardiogram at the late 60th of the last century. The following figure illustrate the historical definitions of MI by the time (Denmark et al., 2019).

There are five major types of MI; the primary coronary event, the second type caused by a problem of oxygen supply and demand, the third type is a sudden cardiac death, the fourth type is Percutaneous coronary intervention and the fifth and the last major type is related to Coronary artery bypass grafting.

Problem statement

Myocardial infarction is a major cause of death and disability worldwide. Predicting the factors that has influences on MI among chronic diseases patients may lead to save lives. Not too many studies conducted to predict MI using data mining techniques. Although it has huge potential to extract hidden knowledge. Many researchers used known bio statistical methods to predict the association between MI and its factors among Patients, but still the data and factors are not available. There is a desperate need to utilize more in data mining techniques in order to use in medical space in Gaza strip. The problem statement can be explained as the following:

- The lack of studies about prediction the MI.
- The health institutions in Gaza Strip and medical staff are not depending on information technology and data mining for medical aspects

Objectives

1.1.1 Main objective

The main objective in this study was to assess the ability of data mining techniques to predict MI among non-communicable patients.

1.1.2 Specific objectives

- To conduct pre-processing operations on the collected dataset.
- To predict MI by using data mining algorithms.
- To evaluate the proposed data mining framework in terms of accuracy.

The importance of the study

Coronary atherosclerosis is a chronic disease with stable and unstable periods leads to MI which is a major cause of death and disability worldwide (Shahwan et al., 2019). Extracting the medical knowledge using data mining techniques and predicting the risk factors for MI among chronic diseases patients in Gaza Strip is going to help the health institutions for putting better plans to save the patients' lives and it is going to alert the patients to check their status.

Scope of the study

The main scope of this research is to use data mining techniques to extract new knowledge from patient's dataset collected for the year 2015 from UNRWA clinics in Gaza Strip.

Research Framework

The research framework consists of four main stages as illustrated in ([Figure 1.1](#)).

Phase 1: Literature review, this phase consists of two main sections. The first section is deep reading to understand and analyze the general topics of data mining. The second sector is reviewing the previous studies and researches related to using data mining in educational purposes and using data mining in prediction, both sectors leads to Expand our knowledge in the problem domain (related works and related techniques used in data mining).

Phase 2: Data collection and preprocessing, the purpose of this phase is to find out a suitable dataset that contains information about students and their evaluations and qualifications. After that, data mining preprocessing techniques such as cleaning, discretization, sampling, feature selection and normalization have been applied. This phase aimed to investigate the data quality and increase the accuracy of the mining.

Phase 3: Data-mining techniques application, after the data set processing, the data mining classification techniques such as Decision Tree, K-Nearest Neighbor, and Naïve Bayesian will be applied. This phase lead to use the suitable techniques.

Phase 4: Performance evaluation, this phase aims to select the best model and evaluate the performance based on testing on actual present cases.



Figure 1.1: Research Framework

Thesis structure

This thesis consists of six main chapters, which are structured around the objectives of the research. The main points discussed throughout the chapters are listed below:

- **Chapter 1: Introduction:** It gives a short introduction about the diabetes disease, the thesis problem and objectives.
- **Chapter 2: Theoretical Foundation:** This chapter presents an overview of data mining techniques and the steps of data mining life cycle process, explains data mining methods, focus on classification methods. Finally, this study focus on Predicting the MI using Data mining techniques.
- **Chapter 3: literature review:** presents other works related to the thesis.
- **Chapter 4: Methods and Experiments:** This chapter explains the proposed approach about using data mining and Predicting the MI using Data mining techniques among chronic diseases patients. An explanation about the data sets used in the experiments,
- **Chapter 5: Results and Discussion:** gives in detail about the sets of experiments, analyzes the experimental results. In addition, it gives a discussion for each experiment and evaluation results.

Chapter 6: Conclusion and Recommendations: discusses the final conclusions and presents possible future works, finally gives some useful recommendations about the thesis.

CHAPTER 2

Theoretical Foundation

CHAPTER 2

Theoretical Foundation

Our research aims to predict MI among chronic diseases patients using data mining methods to extract knowledge of data. Therefore, Current chapter offer the techniques the road of the data mining life cycle phases, processes, explain data mining methods classification methods and continues to explain the MI.

2.1 Data Mining

2.1.1 Introduction

Without doubt. Using computers leads data processing in our society nowadays, which generates large amounts of data in all life aspects such as business, marketing, science, surveillance, economics, fraud detection, sports, medicine etc., Daily, Terabytes of data flow into our computer networks and our storage, the growth of using computerization and the rapid development of data collection and storage tools leads to greater data volume in our community (Pezzè et al., 2012-2013).

The huge amount of data and the rapid growth in the volume of data exceed our human ability to handle the data and to infer the knowledge ([Figure 2.1](#)) without using a powerful tool to extract it. In addition, the accumulated data in the large data repositories is rarely visited. therefore, Mostly, decisions are made based on the data stored in the storage tools, due to the lack of knowledge tools to extract them from the huge amounts of data. (Pezzè et al., 2012-2013).

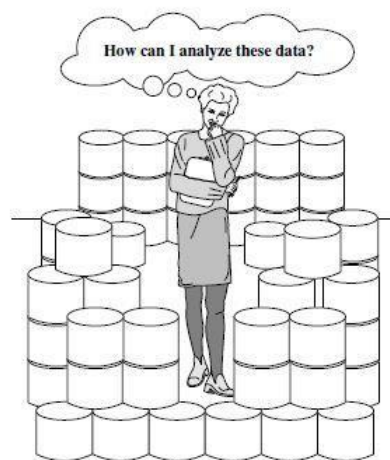


Figure 2.1 The world is full of data and poor knowledge (Han et al., 2012b).

Data mining is known as extracting knowledge from the data (KDD)“Non trivial extraction of implicit previously unknown and potentially useful information about data” (Kolçe & Frasher, 2012). The core step in KDD is Data mining is also defined an analysis of data to find knowledge is an example of that to extract knowledge and arrange knowledge in new and useful ways for the data owner. (Figure 2.2) (Han, Kamber, & Pei, 2012). KDD used by a person who carries out academic or scientific research, databases, Artificial intelligence, pattern recognition, knowledge acquisition, data visualization, and knowledge acquisition for expert systems. The primary goal of exploration and research in raw data is to extract knowledge.



Figure 2.2 : Data mining - searching for knowledge in data (Han et al., 2012)

2.1.2 The KDD Process

Knowledge extraction from databases (KDD) In an automatic way, Big data modeling and analysis. And can say that it falls within the framework of organized operations in order to determine useful patterns of large and complex data.

Data Mining (DM) is the core of the KDD process, this process involves data mining algorithms, Developing the model and discovering unknown knowledge beforehand. The model can be used for analysis, forecasting and understanding of phenomena (Maimon & Rokach, 2009).

The KDD process is iterative and interactive. Basically, KDD process comprises of nine stages as illustrated in figure 2.3, the process is iterative at each stage, which

means that the process implying might need to move back to the previous process to adjust it.

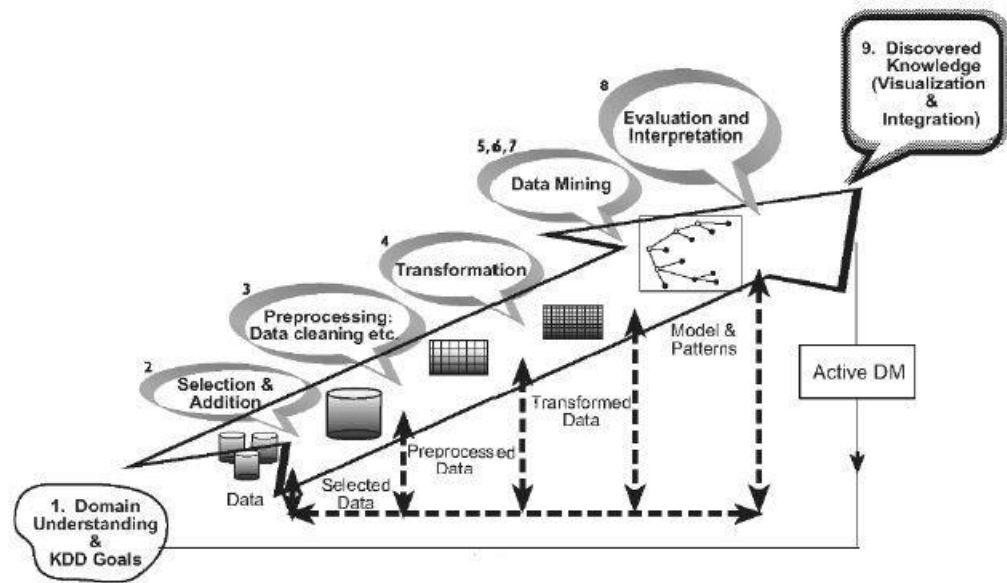


Figure 2.3 :The Process of Knowledge Discovery in Database (Oded & Maimon, 2010).

The excavation process begins with the identification of the target first, and then ends with the narration of knowledge (M. Oded & Lior, 2010). The description of each process explained as following:

1. Domain Understanding and KDD Goals: This process started by understanding the application domain, it is the field of decisions and decision makers must understand the end user's need and the ambience in which the process of discovering knowledge will occur. The people in charge supposed to take in consideration that the KDD process are iterative which means that they might return to this process several times.
2. Selection and Addition: In following step the selected row data set; this step started by determining the data by finding out data availability, obtaining additional Important data, then combining the complete data to discover the knowledge into one data set, including attributes for the process. The importance of this process is very high, because extracting data will depend on the available data to learns and discovers, the study will fail if any data attribute is missed from this step.

3. Preprocessing and cleansing: This stage includes cleaning the data, such as addressing the lack of data and removing the wrong data. Also, these steps use complex statistical tools or data mining tools and algorithms. Without limitation, in the case of a specific feature it is not sufficiently reliable or the attribute missing data, the goal of data mining is to supervised algorithm to predict the missed data.
4. Transformation: the aim of this stage is to find useful features for representing the data, the represented data depends on the action aim, and transformation or to reduce the number of records or the number of fields we can use dimensionality that effective Subject to the representation of static data.
5. Data Mining-Choosing task: This stage depends on the KDD goals and the results of the previous steps. This step may include classification, regression, clustering...etc.
6. Data Mining-Choosing the algorithm: This stage starts by selecting method(s) for searching for patterns among the data. Then, deciding which models and parameters may be appropriate.
7. Data mining-applying the algorithm: This stage includes applying the algorithm iteratively by tuning the algorithm parameters till reaching satisfying results.
8. Evaluation and interpretation: At this stage, the result patterns will be evaluated and interpreted according to the defined goals at the first stage. At this stage, the focus is on understanding the usefulness of the induced model and also documenting knowledge for use in several areas.

Discovered knowledge (Visualization and Integration): This stage is the last stage of the KDD process which determine the effectiveness of the entire process, which comes through the integration the discovered knowledge into another system for further actions.

2.1.3 Data mining functionalities

The variety of data and information repositories leads us to use data mining. It uses data mining functions to determine the type of patterns in these tasks. Generally, it can be said that the mission can be distributing into two categories: predictive and descriptive. Predictive mining tasks perform inference in order to make deduction on the flow data. Descriptive mining tasks are a distinction of general characteristics in any

form of data storage (Singh & Swaroop, 2013). Data mining functionalities are described as below:

Classification: Is one the most important and useful techniques. Classification is useful to address an enormous amount of data, it used predict categorical labels Known as class label (Gorade, Deo, & Purohit, 2017). Among the works that extract models, is also considered from data analysis is what is known as classification. Classified works are classified as supervised. It is used to predict lass labels. There are two known steps for classification; The first step is the learning step, through which the class is created and the second step, which is an important task, which is extraction or predict class labels (Sun & Han, 2012).

Prediction: comparable to classification, prediction of one of the properties of classifications and based on the label ID field predicts a specific value. The prediction is based on lost values and we would like to know them based on previous data. So we can say based on big data and related data that gives a possible value for the new field. (Ramachandran, Liu, Asghar, & Iqbal, 2009; Sun & Han, 2012)

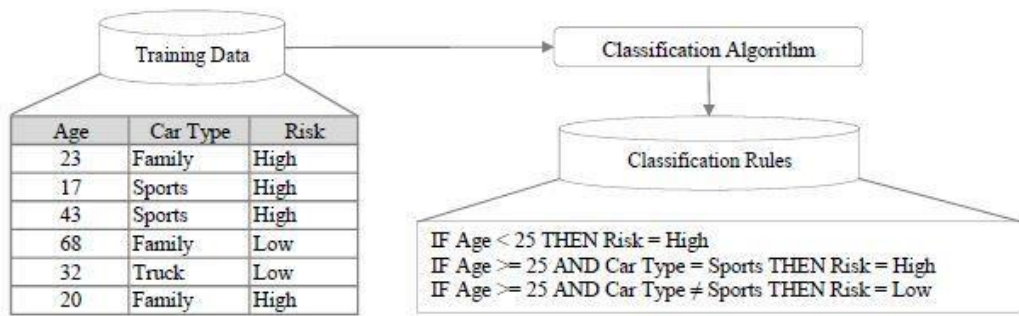
Outlier detection: Outlier detection One of the most complex tasks is the task of extracting data. One of the goals of external disclosure is to discover data and patterns in raw data so that they are not consistent with expectations or their behavior (Gao, et al. 2020). According to (Sun & Han, 2012), outlier detection is also defined as It is a process of discovering incompatible, unexpected, and meaningless data with known and included data, objects are called outliers. In similar words, as if it were generated by a different mechanism. Outliers are different from noisy data. The noise is a different random error in the measurement, which better still to be removed in data mining tasks before outlier detection.

Classification

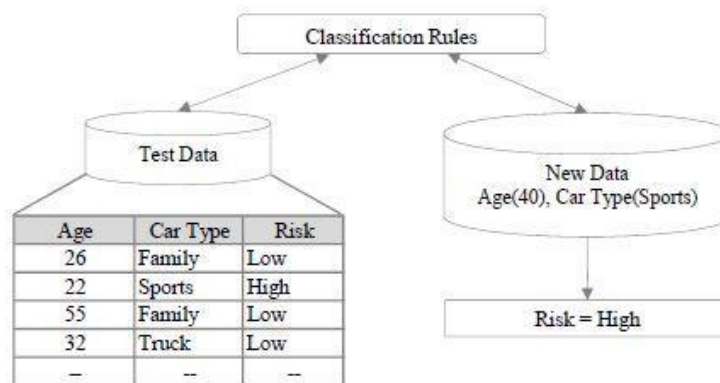
Classification It is the most commonly used knowledge extraction and knowledge mining, and is commonly used in data mining, neural networks, and SVMs. A dozens of quantitative performance measures were proposed with a domination of accuracy, specificity, and sensitivity (Support Vector Machines).

Classifiers are the result of extracting models from the data analysis, it describes the important data classes, also known as supervised classification. It is known that the use of classification is to extract or predict class labels (Sun & Han, 2012).

Data classification process consist of they are two steps. The first is specialized in analysis called learning and analyzes data during training with the help of classification algorithms. (Figure 2.4) illustrate the learning step in data classification process, in this case the class label field is “Risk” that represent whether the insurance transaction is either “High” or “Low”. The classifier represented as classification rules in the figure, the step also can be presented by the function $Y=f(X)$, which means that predicting the Y value by the function of the given data represented by X, by applying the classifier. the step as supervised training because works are taught during the training period, unlike uncensored learning.



(a)



(b)

Figure 2-4 : The Data Classification Process (Han et al., 2012).

The second step is illustrated in [\(Figure 2.4\)](#) is the classification step, where the predictive accuracy of the classifier is assessed. The classifier accuracy is calculated by applying the classification rules, which generated from training data on a given test data (The test data is independent of the training data, which means that the test data were not used to build the classifier), the related class label of each test data is equated with the educated classifier's class prediction for that records. The higher percentage of accuracy means that the higher correctly test data classified by the classifier (M. Oded & Lior, 2010; Sun & Han, 2012) .

2.2 Classification Algorithms

There are many data mining classification algorithms, those algorithms presented over time such as Rule Based classifier, k-Nearest Neighbors and Decision Trees (Mitchell, 1997). This section will discuss a three of these algorithms which has been used in this study.

2.2.1 Naive Bayes

Naïve Bayes is a statistical model that be disposed to work efficiently on text classifications and habitually takes less time for orders of magnitude to train when compared to models like support vector machines. A high degree of accuracy can be obtained using Naïve Bayes model comparing to the current state of the model. Naïve Bayes classification technique based on Bayes Theorem, and it consists of 4 steps, starting by calculating the prior possibility for given class labels, then finding the likelihood probability with each attribute for each class, after that put these value in Bayes formula for calculating the posterior probability, and finally, watching the classes to find which class has a higher probability. Naïve Bayes described by (Borro et al., 2006) and worked as the following:

Let training set D the set of features and their connected class labels. As usual, each feature is represented by an n -dimensional attribute vector, $X = (x_1, x_2, \dots, x_n)$, depicting n measurements made on the tuple from n attributes, respectively, A_1, A_2, \dots, A_n . Suppose that there are m classes, C_1, C_2, \dots, C_m . Given a tuple, X , the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X . That is, the naïve

Bayesian classifier predicts that feature X go to the class Ci if and only if

$$P(C_i|X) > P(C_j|X) \text{ for } 1 \leq j \leq m, j \neq i \quad \dots\dots\dots \text{Equation (2.1)}$$

Thus we maximize $P(C_i|X)$. The class Ci for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad \dots\dots\dots \text{Equation (2.2)}$$

As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \dots = P(C_m)$, and we would therefore maximize $P(X | C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = \frac{|C_i, D|}{|D|}$ is the number of training tuples of class Ci in D.

Given datasets with many attributes, it would be extremely computationally expensive to compute $P(X | C_i)$. In order to reduce computation in evaluating $P(X | C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(X|C_i) = \prod_{k=1}^n P(X_k|C_i) \quad \dots\dots\dots \text{Equation (2.3)}$$

$$P(X|C_i) = P(X_1|C_i) \times P(X_2|C_i) \times \dots \times P(X_n|C_i) \quad \dots\dots\dots \text{Equation (2.4)}$$

The probabilities P can be estimated easily ($P(X_1 | C_i)$, $P(X_2 | C_i)$, $P(X_n | C_i)$) from the training tuples. Recall that here X_k refers to the value of attribute A_k for tuple X. For each attribute, we look at whether the attribute is categorical or continuous-valued. For instance, to compute $P(X | C_i)$, we consider the following:

If A_k is categorical, then $P(X_k | C_i)$ is the number of tuples of class Ci in D having the value x_k for A_k , divided by $|C_i, D|$, the number of tuples of class Ci in D.

If A_k is continuous-valued, then we need to do a bit more work, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation, defined by

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \dots\dots\dots \text{Equation (2.5)}$$

$$P(X_k|C_i) = g(X_k, \mu_{ci}, \sigma_{ci}) \dots\dots\dots \text{Equation (2.6)}$$

To compute μ_{ci} and σ_{ci} , which are the mean (i.e., average) and standard deviation, respectively, of the values of attribute A_k for training tuples of class C_i . We then plug these two quantities into Equation stated at section B, together with X_k , in order to estimate $P(X_k|C_i)$.

In order to predict the class label of X , $P(X_j|C_i)P(C_i)$ is evaluated for each class C_i . The classifier predicts that the class label of tuple X is the class C_i if and only if

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad \text{for } 1 \leq j \leq m, j \neq i \dots\dots\dots \text{Equation (2.7)}$$

In other words, the predicted class label is the class C_i for which (1) (2) is the maximum.

2.2.2 k-Nearest Neighbors

The nearest neighborhood algorithm, or what is called KNN, is subject to supervision, and this process is done based on the closest neighborhood. This is done based on the characteristics and we can say on samples of training. It considers that this closest algorithm is considered slow learning, which means rounding is local approximation and excluding all other calculations until classification. (Phyu, 2009).

To predict the classification of the new object:

- 1) First we determine a user defined parameter “k” which represent number of nearest fields or neighbors.
- 2) We want to measure between the target object and all training samples.

3) The next step is the process of sorting the distance and then determining the closest neighbors based on the closest distance.

4) Hence, I am compiling the classification of the closest neighbors.

5) Then we use the simple, near majority and take it as a value for the new object.

2.2.3 Decision Trees

Among the widely used algorithms is decision tree. Decision tree training algorithms is a classification technique and it has been used in many applications such as manufacturing, medicine, production and financial analysis. The core of decision trees is creating a model which forecasts the rate of a class label based on numerous input variables. As is illustrated in (figure 2.5), a decision tree is a tree-structured plan where each internal node (non-leaf node) represents an attribute, each branch represents a value of the attribute and each leaf node (terminal node) holds a class label (Lior, 2014; Sun & Han, 2012).

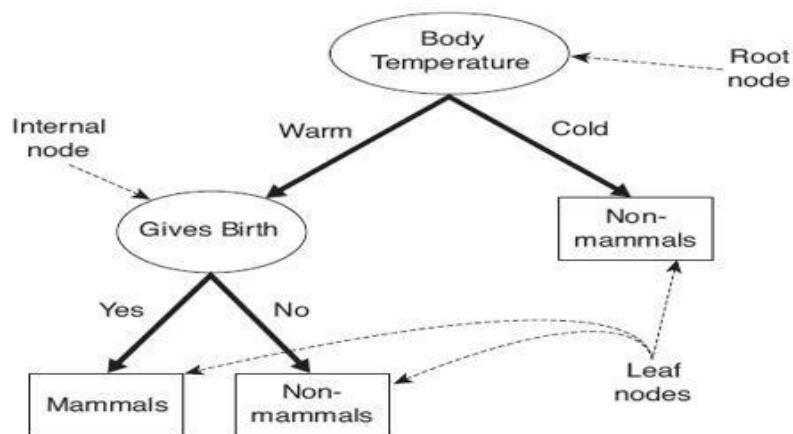


Figure 2-5: Decision tree the mammal classification problem (Tan, 2006)

To understand the decision trees functioning for classification, start by supposing a new object with unknown class label. The new object carried the attribute values and tested against the decision tree. In connection with the work of this algorithm, that is, it works to track data from top to bottom. The process begins with tree tracking, but with relevant data. The training group is divided into small groups,

and this is frequently done into smaller groups, and the decision tree is summarized by (Kolçe & Frasherri, 2012) as tracks:

Calling the algorithm with three parameters, the first parameter is "D" which illustrate a dataset (initially, it consists of a complete set of training data and their related class labels); second parameter is the attribute list, which means a list of attributes describing the data; and the last one is the attribute selection method, which specify indicative procedure for selecting the attribute that “best” classify the given data according to class, this procedure uses a criterion selection measure such as information gain.

Create a single node “N” which figures to the training data in “D”.

Checking the data in "D" if all of the same class, the "N" becomes a leaf and labeled with that class.

Else, the algorithm calls criterion selection method to determine the node attribute “N” by determining the “best” way to divide the data in “D” into individual classes, also the algorithm choose which branches to grow from node “N”.

The “N” node is labeled with the splitting criterion. The "N" node will produce a branch is grown from it for each outcome of the splitting criterion. A division of the data "D" will be operated accordingly.

The same process is frequently used as the decision tree to form the tree from each resulting section, “D_j”, of “D”.

$$InformationGain(a_i, S) = Entropy(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \cdot Entropy(y, \sigma_{a_i=v_{i,j}} S) \dots\dots\dots Equation (2.8)$$

where:

$$Entropy(y, S) = \sum_{c_j \in dom(y)} - \frac{|\sigma_{y=c_j} S|}{|S|} \cdot \log_2 \frac{|\sigma_{y=c_j} S|}{|S|} \dots\dots\dots Equation (2.9)$$

The recursive division operation terminates if any of the following conditions is true: All the data in partition “D” belong to the same class; and there are no remaining

attributes on any data that may be further partitioned. In this case, the node “N” converts into a leaf and labeling it with the most common class in “D”; There are no data for a given branch will lead to “Dj” empty. In this case, a leaf is created with the majority class in D.

The decision tree resulting is returned. To determine which attribute will be chosen to divide the data when creating a decision tree, an attribute selection measure. This process provides for each field a specific classification, and therefore the best classification is chosen and the tree node is classified with the letter D under a specific classification criterion. The algorithm works through it and then the branches are planted for each result of the criterion, and the data is divided according to this step or divisions. Information gain is one of the popular attribute selection measures. This process, I mean division, takes place under the standard of dispensing impurities, as the impurity scale is used as a measure of this process (M. Oded & Lior, 2010).

2.3 Myocardial Infarction

Nowadays, the term (MI) It is clear from what it was that, based on research and investigating information from more than one site, it was found that heart attacks have claimed the lives of many people who have been exposed to it. (Kristian Thygesen et al., 2007). It turned out that in the later years of the period, it became clear that a relationship must be established between the blockage of the arteries of the heart and (MI). Still, the first clinical descriptions appear in the beginning of the 20th century as a There is no doubt that the relationship between this blockage and the coronary stroke of the cardiovascular system. Regardless of what we noticed is that a long time has passed until they found this result in the event of an autopsy through a study that showed that there was no blood clot in the arteries and the percentage was 31% of patients who died from sperm from those who were suffering from the disease (MI) over time. By time, several MI definitions have been used, which leads to controversy and confusion. Till the for an electrocardiogram, the WHO developed a definition for this electrocardiogram at the late 60th of the last century. The following (Figure 2.6) illustrate the historical definitions of MI by the (Kristian Thygesen et al., 2019).

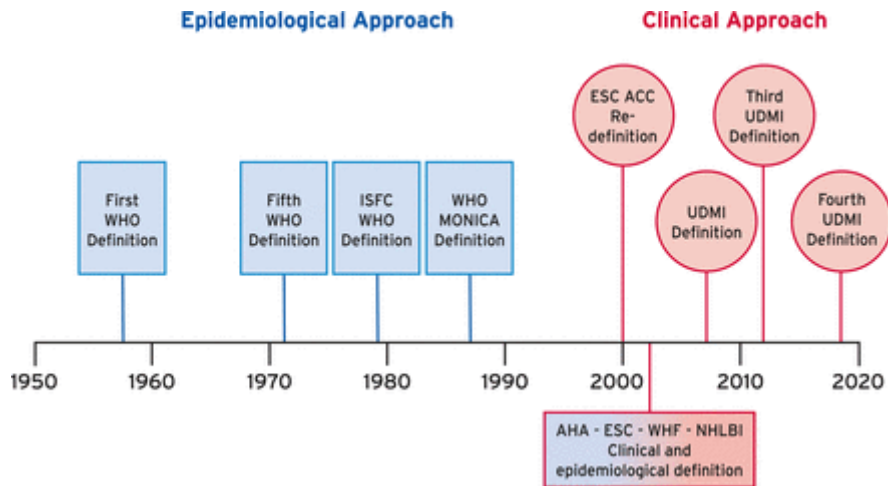


Figure 2-6: History of Documents on the Definition of Myocardial Infarction

Pathologically, MI is defined as "myocardial cell death due to prolonged ischemia. Diminished cellular glycogen, and relaxed myofibrils and sarcolemma l disrupt

ion, are the first ultra-structural changes and are seen as early as 10 to 15 min after the onset of ischemia" (Jennings & Ganote, 1974).

2.3.1 Major Types of MI

According to (ZOLER, 2006) there are five major types of MI. We can formulate it as follows:

- The first type is the primary coronary event, which related to plaque rupture or dissection.
- The second type caused by that any defect in the arrival of the element known to oxygen or even the possibility of taking or exiting it, or in the event of coronary obstruction or in the event of an irregular heartbeat or a drop in blood pressure or a vertebra.
- The third type is a what is known as sudden death which include signs of a lack of nourishment in the heart or lack of fluid to reach it.
- The fourth type is percutaneous coronary intervention.
- The fifth and the last major type is related to coronary artery bypass grafting.

2.3.2 Symptoms of MI

Because of many possible conditions of MI, there are many MI symptoms and numerous. Some of MI symptoms are chest pain, sweating, nausea or vomiting, and fainting, shortness of breath, weakness, fatigue and also it may occur without symptoms and called silent infarction (Braunwald et al., 2001). In details the symptoms divided as follow to three categories:

Chest pain: it is one of the most common MI symptoms, it turned out that there are symptoms that there is a shortage of breathing and that these symptoms are accompanied by pain in the neck and may reach the left hand and perhaps to the right hand and back, as well as the lower jaw, but doctors rely on diagnosing the upper abdomen as a first stage (Anderson et al, 2013).

Associated symptoms: some symptoms are associated with chest pain, but it also may occur without chest pain, the associated symptoms such as sweating, nausea or vomiting, and fainting (K Thygesen, Alpert, & Jaffe, 2013). Other symptoms occur for women such as shortness of breath, weakness, and fatigue (Canto, J. et al 2007).

Silent infarction: according to (Valensi, Lorgis, & Cottin, 2011). The silent infraction is the MI which happens without any symptoms at all. At this case the MI cases discovered later after the person death using blood enzyme tests, or at autopsy.

2.3.3 MI diagnosis

The diagnosis of MI begins with finding the patient medical history to identify if disease has occurred in the past, and continues with the First of all, there is a physical examination that doctors rely on and accompanies this known electrical examination of the heart. The physical examination depends on the symptoms appears on the patient, while the main reason to use In order to detect ischemia or heart attack known as (MI), an electrocardiogram is used (Kasper et al. 2005).

Using ECG (also referred to as an EKG or electrocardiogram) measures the heart electrical activity. The activity levels changes in the heart can signal a heart disease(Cannon et al. 1999). The cardiac markers are the leaked out proteins from the

injured myocardial cells. The enzymes can be diagnosed through the blood tests to determine if there is any MI (Apple et al., 2005).

2.4 Summary

In this chapter, four main categories were presented of theoretical foundation in this thesis, data mining, classification, classification algorithms and myocardial infraction.

In data mining, the main aspects and data mining functionalities were represented, in classification, a general review of classification in data mining were presented, brief description for some algorithms use classification techniques were shown, and finally a brief information about myocardial infraction were represented as the core of this research to be predicted in this research using data mining.

CHAPTER 3

literature review

CHAPTER 3

literature review

In this chapter, previous related works reviewed. This chapter combined of three parts, the chapter starts by the introduction, then review of traditional risk factors of MI e.g. age, sex, smoking history, family history, lipid profile, obesity, physical inactivity, hypertension, diabetes and finally using data mining in MI prediction.

3.1 Introduction

With no differences among many chronic diseases, MI requires multispecialty care, more self-care is needed from patients such as blood-sugar, adherence to medication, diet, exercise, and recommendations. Huge number of works related to chronic diseases and the use of data mining for diagnosis and predicting have motivated this research. Practicality, low-cost communication is needed to address the data requirements of any health system.

3.2 Cardio Vascular Risk Factors

Traditional cardiac risk factors prevalent among Palestinians females in Gaza with common combination of these risk factors. The prevalence of CAD in among women in Gaza population was about 50% compared to normal coronaries. Females with CAD have high distribution of Hypertension, with a double risk (Abed & Jamee, 2015). However, Diabetes Mellitus is a popular major risk factor of CAD among Palestinians females in Gaza (Jamee & Abed, 2014). The presence of diabetes in women increasing CAD risk by three to seven times in comparison to double or triple increased risk among diabetic men (Huxley, Barzi, & Woodward, 2006). In Gaza, diabetes was double fold higher among females with CAD in comparison with healthy coronaries (Abed & Jamee, 2015).

3.2.1 Age

Age is a significant associated factor with functional changes in skeletal muscle. These changes able to decrease energy consumption and needed activity, this rise the prevalence of morbid obesity, which surly contributed to the increase prevalence of

diabetic, increased lipid blood level and arterial hypertension in more than forty years old individuals. However, in very old men and women total cholesterol and subtotal cholesterols in addition to triglycerides noted to be lower (Marín-García & Goldenthal, 2008).

Cardiovascular risk is predisposed by age but it depends also on other risk factors (Schnabel et al., 2009). Long-term burden of cardiovascular risk factors increase the need of remodeling of these cardiovascular factors (Kannel & Vasan, 2009).

3.2.2 Sex

Risk factors in women can be sex-specific, usually related to sex-hormones, or conditions which are more prevalent in women condition. Reproductive and pregnancy-related history is recommended to elucidate relevant risk factors (Brown, Bayley, Harrison, & Martin, 2013). Pregnancy-related conditions including gestational hypertension, preeclampsia and gestational diabetes are associated with a 2-fold increased risk of myocardial infarction and brain stroke (Bellamy, Casas, Hingorani, & Williams, 2007; Fadl Elmula et al., 2014; Kessous, Shoham-Vardi, Pariente, Sherf, & Sheiner, 2013). Autoimmune diseases such as rheumatoid arthritis and systemic lupus erythematosus disproportionately affect women and confer an increased risk of cardiovascular disease (Mason & Libby, 2015).

3.2.3 Smoking

Internationally, it is predicted that 1.6 milliard people will smoke cigarettes (Longo et al., 2012). Approximately one among five adult persons are currently smoke with a greater proportion in men than women as the example in America (Go et al., 2014);(Mosca, Barrett-Connor, & Kass Wenger, 2011). Cigarettes smoking is well known to have a major effect on body health particularly cardiovascular system. Smoking is considered as the most important preventable risk factor of CVD (Go et al., 2014); (Anderson et al., 2013). Smoking cessation significantly reducing the risk of CVD and CHD morbidity and mortality (Detection & Adults, 2002). Overall mortality among smokers is three times higher than non-smokers. According to the expectation for 2030, the burden of disease associated with tobacco smoking will lead to around ten million deaths every twelve months, from around 5 million deaths caused by tobacco, nearly one and half million are CVD patients (Longo et al., 2012) (Go et al., 2014).

Smoking is a cardiovascular risk factor not only for smokers but also for those negative smokers by exposure at home or at work to smoking, they also called second hand smokers and the risk of developing coronary heart disease among them increases by 25% to 30%. Regardless of the measures taken to raise awareness about the believed bad effect of cigarettes smoking on health, it still remains a dominant risk factor for cardiovascular disease. (Longo et al., 2012) .

3.2.4 Family history of cardiovascular disease.

Positive family history of the first-degree relative (parent, sibling, or offspring) is a known risk factor of coronary heart disease, and it is good to predict and identify younger patients (Kones, 2011); (Detection & Adults, 2002).

A familial history of premature heart attack increase the risk of CVD two times more in men women (Mozaffarian, Benjamin, Go, Arnett, Blaha, Cushman, Das, et al., 2015). Sibling history of CVD increases the risk of cardiovascular disease by about 50% in males compared to females (Mozaffarian, Benjamin, Go, Arnett, Blaha, Cushman, Das, et al., 2015).

3.2.5 Lipids

Cholesterol is a lipid substance that lipoproteins are responsible for its traveling in the blood (Detection & Adults, 2002). The current AHA guidelines recommend total cholesterol of more than two hundred mg/dL for healthy individual may reduce heart and brain stroke attacks risk (Go et al., 2014). Low-density cholesterol (LDL) covers around 65% percent of the total serum cholesterol and is the major anthropogenic lipoprotein (Grundy et al., 2004).

3.2.6 Triglycerides

An overnight fasting triglyceride level averaged 130.3 mg/dL and level of equal or more than 150mg/dL is considered a risk factor for stroke either in heart or brain vessels. Overall, mean level of triglycerides among males is higher than females (McGowan & Keet, 2013; Mozaffarian, Benjamin, Go, Arnett, Blaha, Cushman, Das, et al., 2015).

3.2.7 Low density lipoprotein

Adult Treatment Panel III was classified the LDL cholesterol level as borderline high (130 -159 mg/dL), high (160 - 189 mg/dL); and very high (190mg/dL and above). Data from the National Health and Nutritional Examination Survey (McGowan & Keet, 2013) indicated that the mean adults LDL cholesterol is about 115.8mg/dL was accompanied with other comorbidities as cardiovascular disease, hypertension and diabetes mellitus.

3.2.8 High density lipoprotein

High-density lipoprotein involves about 25% of the total serum cholesterol (Detection & Adults, 2002). A HDL cholesterol level of around 50 mg/dL in adult men and women is as a risk factor for CVD (McGowan & Keet, 2013). The NHANES 2007-2010 data indicated a mean HDL level around 50mg/dL with greater racial and gender variation in HDL levels in comparison to LDL levels. The triad of elevated triglycerides, LDL and decreased HDL are strongly related to diabetes (Payne, 2012) .

3.2.9 Obesity

Obesity is well-defined as increased body mass index (BMI) of more than 30 Kg/m². Obesity is a well-known risk factor of CVD, HTN, hyperlipidemia, and DM (Go et al., 2014);(Mozaffarian, Benjamin, Go, Arnett, Blaha, Cushman, Das, et al., 2015). A complete CVD risk assessment includes measures of BMI and for central adiposity. The health risks by obesity may also be mediated by hyperlipidemia hypertension, and diabetes mellitus (Longo et al., 2012);(Detection & Adults, 2002). Central adiposity (abdominal fat) is believed to have significant role in atherosclerosis compared to total body fat (Everson-Rose et al., 2009).

3.2.10 Physical inactivity

A major risk factor the development of CVD and diabetes mellitus is the physical inactivity (Anderson et al., 2013). The AHA guidelines specified a 150 minutes a week of moderate to intensity activity are needed to achieve health heart (Go et al., 2014). According to 2014 National Health Interview Survey data, the inactivity gender difference was clear since females were less active than males and increased

with age (Mozaffarian, Benjamin, Go, Arnett, Blaha, Cushman, De Ferranti, et al., 2015).

The mortality rate was decreased by 27% which was associated with adherence to the recommended CVD risk factor guidelines in individuals without chronic conditions such as diabetes mellitus, cancer, myocardial infarction, angina, cardiovascular disease, stroke, or respiratory disease. However it reduced by up to 46% among individuals with chronic comorbidities mortality (Schoenborn & Stommel, 2011). Physical activity has been reported to reduce all type of lipids levels in blood. Moreover, it improved insulin sensitivity, and hypertension status (Detection & Adults, 2002).

3.2.11 Hypertension.

Hypertension (HTN) is a medical term used to describe high blood pressure, is known to be a popular and a preventable risk factor for CVD and brain stroke. High blood pressure lead to extensive force of the blood against arterial walls which creates lesion to internal wall of the artery results in scar tissue and by time would lead to narrowing of the artery diameter (Go et al., 2014). Seven or eight of ten persons who experience their first heart attack or first brain stroke have HTN (Go et al., 2014).

Unfortunately, the treatment of HTN is unable to reduce the coronary heart disease risk (Grundy et al., 2004). It estimated that 78 million (one third of whole population) adults in the USA above 20 years of age are HTN. Hypertension is more prevalent in men 45 years and younger compared to women. However, women are about as likely as men to develop high blood pressure during their lifetimes (Go et al., 2014; Mozaffarian, Benjamin, Go, Arnett, Blaha, Cushman, De Ferranti, et al., 2015). However, high blood pressure affects more women that are 65 years and older compared to men of the same age category (Go et al., 2014; Mosca et al., 2011).

3.2.12 Diabetes

Diabetes mellitus (DM was defined by WHO to have fast blood sugar more than 110 mg/dL. The National Health and Nutritional Examination Survey (NHANES) 2012 statistics exposed that 19.7 million people are diabetic with confirmed diagnosis and about 50%% of adults met the criteria for type 2 diabetes mellitus (Mozaffarian, Benjamin, Go, Arnett, Blaha, Cushman, Das, et al., 2015). The increased incidence of

diabetes mellitus (DM) has increased the incidence of CHD, brain stroke, and heart failure (Go et al., 2014). Diabetics alone is considered at a high risk for CVD regardless of the presence of other risk factors (Detection & Adults, 2002; Payne, 2012). Moreover, diabetes is associated with increasing of the prevalence of elevated LDL, HTN, and obesity (Go et al., 2014).

3.3 MI and BMI with traditional statistical method:

A study was held by (Wolk, Berger, Lennon, Brilakis, & Somers, 2003) to investigate the risk Factor for Unstable Angina and (MI), the study runs on patients with coronary artery disease. The study sample was 502 patients undergoing coronary angiography, the results of this study showed that, patients with established coronary atherosclerosis, BMI, as well as CRP and number of coronary lesions, are independently associated with acute coronary syndromes. Also, the results proofed that the risk increased even at mildly elevated BMI levels.

Another case study, was held to find out the Obesity the risk of (MI), the study was hell on 27000 patients from 52 countries. The results of this study showed that waist-to-hip ratio shows a graded and highly significant relation with MI risk worldwide. The study find out that the redefinition of the obesity based on waist-to-hip ratio instead of BMI the increases the estimate of (MI) attributable to obesity in most ethnic groups (Yusuf, S., et al. 2005).

In Brazil,(Silveira, Vieira, Jardim, & Souza, 2016) a study aimed to investigate the prevalence of obesity and associated factors. In elder people, the study emphasis on the occurrence of other diseases and on food consumption. Cross-sectional sampling study performed with 418 elderly patients, the results showed that Obesity had a high prevalence in the evaluated elderly population and was associated with food consumption, musculoskeletal disease, diabetes mellitus, and acute (MI).

3.4 Diabetes and Data Mining:

In India,(Sankaranarayanan & Perumal, 2014) applied a Rule Set and Decision Trees classification methods on a dataset. The dataset created by George John from a data repository at <http://kddics.uci.edu> to determine if the person have diabetes mellitus or not. The study used data mining classification methods to diagnose new cases, if the cases results where positive to diabetes or not. The difference among the research and

this study, is in the context of the discovering results, where our research will discover if BMI could be a factor that has influence on MI among diabetes or not.

Another study was held by (Brown et al., 2013). The study aimed to help explain suggested doses for diabetes patients, Euclidean distance has been used to determine the distance between numeric values, the cases which returned from the knowledge base with the closest similarity to the query have been retrieved. At first, the query is defined by asking the user a series of questions to build a picture of the new situation. Once the query is ready, the new problem has been defined and the similarity comparisons were run against the knowledge base. The context differences among this study and ours is our prediction association between BMI and MI among diabetics.

3.5 MI with data mining

Huge number of researches in health field, especially in heart diseased field. The use of data mining developed predicting module and analytical studies to identify those at risk is minimizing the disease risk factors (Safdari et al., 2013).

Performing Data analysis by using SPSS and Clementine version 12. There are Seven predictive algorithms and one algorithm-based model for predicting association rules were applied on the data. Accuracy, precision, sensitivity, specificity, as well as positive and negative predictive values were determined and the final model was obtained. Based on the association rules, five parameters, including hypertension, DLP, tobacco smoking, diabetes, and A+ blood group, were the most critical risk factors of (MI). Among the models, the neural network model was found to have the highest sensitivity, indicating its ability to successfully diagnose the disease Risk prediction models have great potentials in facilitating the management of a patient with a specific disease. Therefore, the changing on life style and the health interventions can be conducted based on these models to improve the health conditions and minimize the risk on patients. The difference among our study and this study is the use of different factors which used to predict the association rules for MI.

Another study conducted by (Xing, Wang, & Zhao, 2007) aimed to develop data mining algorithms for predicting survival of CHD patients, this study based on sample of 1000 patient's cases size. To include 1000 CHD cases, the study carried out a clinical observation and a 6-month follow up. Each survival case information' is gathered via follow up. Based on the data, three popular data mining algorithm were applied to develop the prediction models using the 502 cases. For comparison purposes, 10-fold

cross-validation methods is used to measure the unbiased estimated of the three prediction model. The results came out as following: The SVM is the best predictor with **92.1** % accuracy on the holdout sample, while artificial neural networks came out to be the second with 91.0% accuracy and the decision tress models came out at the third place with 89.6% accuracy.

The following table is compare between the main articles which was closely related to this work in term of the aim of the study, the used methods and the appeared results.

Table 3. 1 Comparison between the main article related to this work

#	Authors	Year	Aim	Method	Results
1	Nag, Procheta, et al (Nag, Mondal, Ahmed, More, & Raihan, 2017).	2017	They predicted the risk of heart attack for a person complaining of chest pain or equivalent symptoms.	They used c4.5 decision tree algorithm, which its accuracy improved by random forest, we used decision tree, KNN, and Naïve Bayes algorithms	KNN was the highest accuracy.
2	(Safdari et al., 2013)	2013	They predicted the risk of heart attack by data mining except.	They used data mining algorithms with multi model (C5, C&RT, QUEST, CHAD).	Accuracy score for (C5, C&RT, QUEST, CHAD) were (85.7%, 79.1%, 72.28%, 93.4%) respectively. While, in our study we used data mining algorithms by the decision tree model which gained accuracy of 92.54%. In addition, we applied KNN and Naïve Bayes algorithms as a sensitivity tests which gave similar accuracy of

					(95.46%, 93%) respectively. Therefore, we conclude that KNN was an applicable algorithm for clinical purposes with the highest precision.
3	Priyanka et al (Priyanka, N., & RaviKumar, P. 2017).	2017	Heart diseases prediction by Naïve Bayes	They used the Naïve Bayes for heart diseases prediction.	Naïve Bayes for heart diseases prediction scored 82.3% accuracy.
4	Hasan et al (Hasan, Mamun, Uddin, & Hossain, 2018).	2018	Heart diseases prediction by different classification techniques by using info gain feature selection technique and removing unnecessary features.	Different classification techniques such that KNN, Decision Tree (ID3), Gaussian Naïve Bayes, Logistic Regression and Random Forest are used on heart disease dataset for better prediction.	They concluded that Logistic Regression performed better (with a classification accuracy of 92.76%) among all approaches

3.6 Summary

As presented in previous sections, we reviewed the known risk factors of MI by many epidemiological studies either by biostatistics science or by data mining. We used data mining methods to extract knowledge from patient's data which provided by UNRWA clinics in Gaza strip. In contrast, our proposed method benefits from the results of using the data mining methods to monitor and predict the MI among patients.

CHAPTER 4

Methods and Experiments

CHAPTER 4

Methods and Experiments

This chapter aims to explain the chosen approach about using data mining to predict MI for chronic diseases patients. The study attempts to construct prediction data mining framework using Decision Tree, Naïve Bayes and KNN. To implement and evaluate this research, various steps have been performed was shown on [\(Figure 4.1\)](#).

First of all, data set must be understood and prepared. The data set has been gathered from the UNRWA clinics. After that, implementing data mining methods to build a module for prediction MI from the data set. Then, evaluating the data mining system.

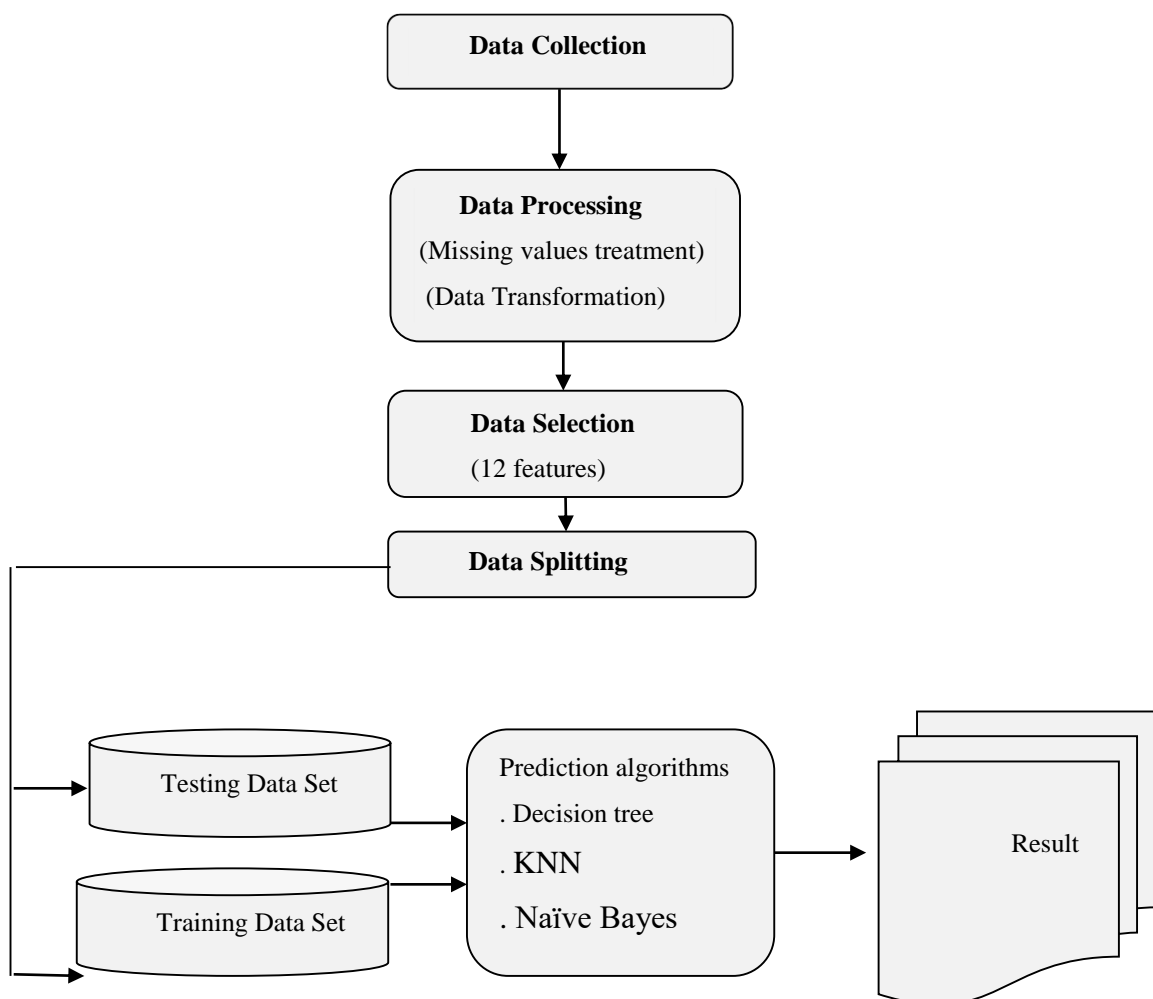


Figure 4-1: Prediction Model Using Data Mining Techniques

This chapter is organized as follow:

- Section 4.1 Presents the taken steps to understand and prepare the data set for implementing data mining methods.
- Section 4.2 Experiments Setup.
- Section 4.3 Data pre-processing.
- Section 4.4 Describes the data mining methods that have been used to build the MI prediction module from the data set.
- Section 4.5 Discusses evaluation the data mining system.
- Section 4.6 Is the summary of whole chapter.

4.1 Collect and preparing the data set

The chronic diseases data were obtained from the UNRWA database for the year 2015 in one excel file for each clinic of the 5 clinics as shown in the [\(Table 4.1\)](#).

4.1.1 Data set attributes

The chronic diseases patients had a hard attack (MI) ' instance counts in 5 of UNRWA clinics distributed in Gaza strip listed in [\(Table 4.1\)](#). Al Remal clinic with 235 patients, then Khan Younis with 96, Al Nusairat clinic counts 205, Al Saftawi clinic with 137 and Al Shabora clinic counts 112. All of them count 785.

Table 4-1 :Chronic diseases patients instant count

#	Clinic Name	No. of instances
1	Al Remal	235
2	Al Saftawi	137
3	Al Nusairat	205
4	Khan Younis	96
5	Al Shabora	112

The file contains patient's personal data, health data, and blood glucose measurements.

[\(Table 4.2\)](#) shows the attributed risk factors associated significantly with CAD derived from literatures along more than two previous decades (see related work chapter three)

e.g. age, gender, smoking status, obesity, history of hypertension, history of diabetes, lipid profile, previous family history.

Table 4-2:Chronic diseases patient’s data set attributes

	Attribute Name	Description
1.	Patient age	Age of patient in years from date of birth If patient age <= 45 years, Age value = 0 If patient age 46 – 55 years, Age value = 1 If patient age 56 – 64 years, Age value = 2 If patient age >= 65 years, Age value = 3
2.	Patient Gender	Patient Gender (Male, Female)
3.	Blood Group	Patient Blood Group (A-, A+, AB-, AB+ ... etc.)
4.	Diabetes Mellitus (DM)	If Diabetes Mellitus Controlled with lifestyle modification, DM value = 0 If Diabetes Mellitus Controlled with medications, DM value = 1 If Diabetes Mellitus Uncontrolled with medications, DM value = 2 If Diabetes Mellitus Diabetes Mellitus with Proteinuria, DM value = 3
5.	Smoking Score	One of the risk factors, No, Smoking Score value = 0 Yes, Smoking Score value = 3
6.	Cholesterol	If Cholesterol < 160 mg/dl, LD value = 0 If Cholesterol 160 - 199 mg/dl, LD value = 1 If Cholesterol 200 - 249 mg/dl, LD value = 2 If Cholesterol ≥ 250 mg/dl, LD value = 3
7.	BMI Obesity Score	Body Mass Index If BMI ≤ 29, Score BMI value = 0 If BMI 30 – 34, Score BMI value = 1 If BMI ≥ 35, Score BMI value = 2
8.	Family History	Family History (Close relatives with hereditary link (parents, brothers, sisters, uncles and aunts) with cardiovascular diseases including strokes, (MI) and hypertension) One of the risk factors, Negative, Family History risk factor value = 0 Positive, Family History risk factor value = 2
9.	FBG	Fasting blood glucose measurement

[Table 4.3](#) shows the attributed complications which had been excluded from the prediction data mining framework of the study. This exclusion was intended to increase

the validity of prediction since these complications are proved to increase the incidence of MI.

Table 4-3: Excluded Factors

#	Factor Name	Description
1	Previous MI	Previous attack of myocardial infarction
2	CHF	Congestive Heart Failure
3	Stroke	Previous attach of cerebrovascular accident
4	ESR D	End-Stage Renal Disease
5	Blindness	Los of vision
6	Amputation	Loss of an organ or a part of an organ

4.2 Experiments Setup:

In this section, the experimental environment is described, and the used tools for the experiments are determined. Finally, the experiments settings of the research are determined.

4.2.1 Experimental Environment:

The experiments were applied on a machine with the following properties: Intel(R) Core(TM) i7-2670QM CPU @ 2.20GHz, 8.00 GB RAM, 500 GB hard disk drive and Windows 10, 64-bit operating system installed.

4.2.2 Experimental Tools:

The used tools for the experiment are the following:

1. **Rapid Miner program:** The Rapid Miner program is a software platform, this software is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including results visualization, validation and optimization.
2. **Microsoft Office Excel:** The Microsoft Office excel is used for organizing and storing the datasets in tables and to perform some simple preprocessing.

4.3 Data preprocessing

For data mining, data preprocessing has been defined as " the set of techniques used prior to the application of a data mining method" (García, Ramírez-Gallego, Luengo, Benítez, & Herrera, 2016). Usually, Real-world data is imperfect, including inconsistencies and redundancies, the data couldn't be applied directly for data mining process. So, Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing (Wu, 2012).

4.3.1 Missing values

Missing data are caused by various reasons, such as the data was not available or data have been missed during the data entry procedures. The missed data filling or removing is very important operation to ensure that acceptable results are given. Rows removal is usually done when many attributes are missing from the row. Ignore records with missing values. This method is based on ignoring cases with missing attribute values. This method is not effective, unless the record contains several attributes with missing values. This is usually done when the class label is missing

As shown in [\(Figure 4.2\)](#).

F	G	H
PATIENTDOB	MI	OCCUPATION
1/1/1945	0	Employee - موظف/ة
1/1/1945	0	Employee - موظف/ة
1/1/1945	1	Employee - موظف/ة
1/1/1945	0	UN Employed - لا يعمل
1/1/1945	?	Homemaker - ربة بيت
1/1/1945	?	UN Employed - لا يعمل
1/1/1945	?	UN Employed - لا يعمل
1/1/1945	0	UN Employed - لا يعمل
1/1/1945	0	Homemaker - ربة بيت
1/1/1945	0	Homemaker - ربة بيت
15/3/1959	0	Homemaker - ربة بيت

Figure 4.2: Show The Missing Value.

In this thesis experiment, the missed values Use Global constant to fill in missing value: Replace all missing values by same constant such as "unknown". Although this method is simple but it is not recommended because results with "unknown values are not "interesting". ("CHF", "Stroke", "ESRF", "Blindness",

“Amputation”, “EBP”, “DM”, “Score Smoking”, “LD”, “Score BMI”, “Inactivity”, “Family History”, “Result1”).

Ass shown in [\(Figure 4.3\)](#).

K	L	M	N	O
VISIT_CATEGORY	VISIT_TYPE	VISIT_DATE	BMI	PHYSICAL_INACTIVITY
NCD	c	19/1/2015	32.46	0
NCD	c	19/1/2015	unknown	unknown
NCD	c	unknown	32.46	0
NCD	c	19/1/2015	32.46	0
unknown	c	19/1/2015	32.46	0
NCD	c	19/1/2015	32.46	0
NCD	c	19/1/2015	32.46	0
NCD	c	19/1/2015	unknown	0
NCD	c	19/1/2015	32.46	0
NCD	c	19/1/2015	32.46	0

Figure 4.3: Show Unknown Value.

4.3.2 Data Reduction

The techniques of data reduction can be applied for obtaining a reduced representation of the data set, attribute subset selection is one of data reduction techniques. Attribute subset selection aim at selecting a subset of the attributes or features, which describe the data in order to obtain a more essential and compact representation of the available information.

In this thesis experiment, we use feature selection to choose attributes are “Patient Gender”, “Age”, “EBP”, “DM”, “Score Smoking”, “LD”, “Score BMI”, “Inactivity”, “Family History” and “Result1”, because these attributes used for building a classifier to predict MI. Selecting Most Relevant Fields.

We used the rapid miner properties to implement this feature (See Figure 4.4).

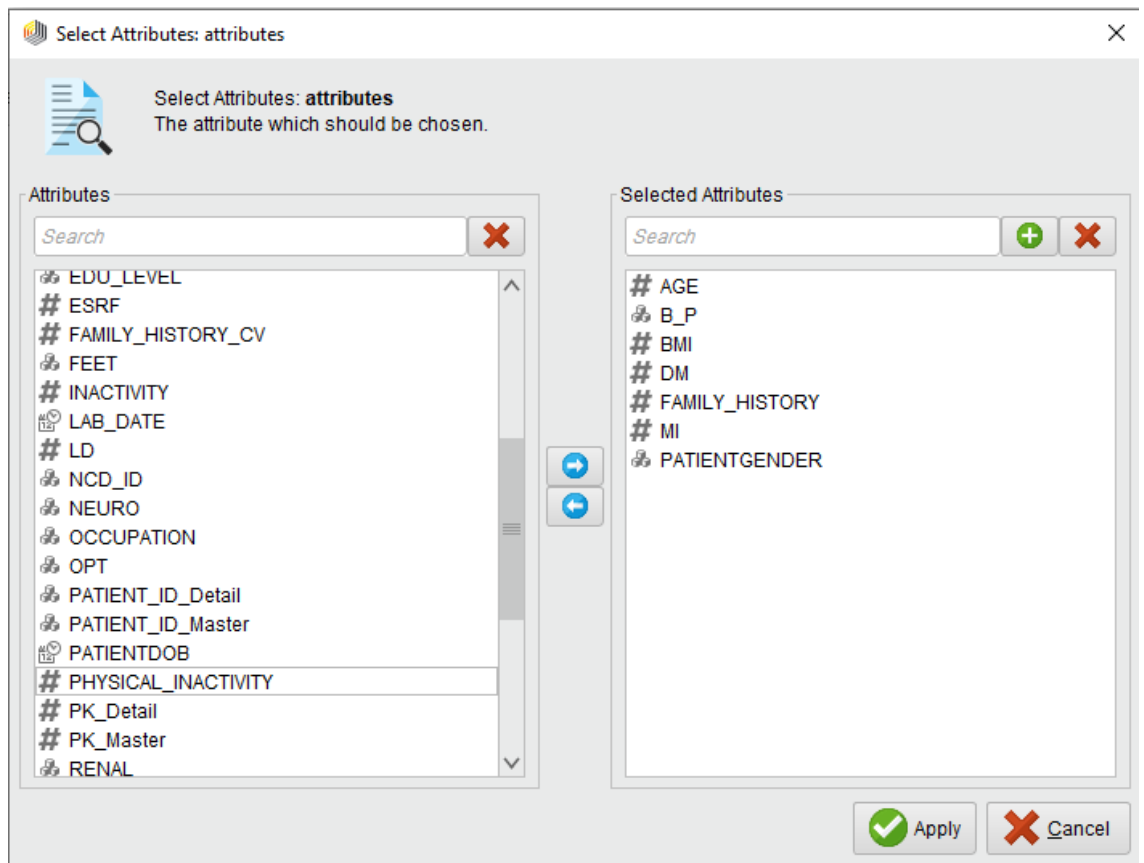


Figure 4.4: Features Selection

4.3.3 Data Transformation

In data transformation, the data are transformed or consolidated into forms appropriate for mining, data discretization is a form of data transformation. Discretization is replacing a numeric attribute values (e.g., age) by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior).

In this study, the following attributes were discretized:

- “Patient DOB” (Patient date of birth): after represents the date values as the year the patients were born:
 - Children: With upper limit 2016 and give it value 1.
 - Young: With upper limit 2000 and give it value 2.
 - Old: With upper limit 1976 and give it value 3.

4.4 Data mining methods

The main part in this research approach is the data mining methods, it is used for MI Prediction. Based on several factors, such as patient age, lifestyle, smoking, blood glucose measurements and others. In this research, the classification method has been chosen for analyzing the MI patient's data and extracts a model to clarify the set of cumulative risk factors that lead to of MI occurrence.

The steps of preparing the data and creating the classifiers is explained below for each late prediction which may occur to the patient by applying a decision tree classification method.

4.4.1 Decision Tree

The following process (Figure 4.5) illustrated the main process of the decision tree method. Decision tree method applied on the dataset to build a classifier for predicting if the patient will be infected with (MI), this method is implemented via Rapid Miner tool:

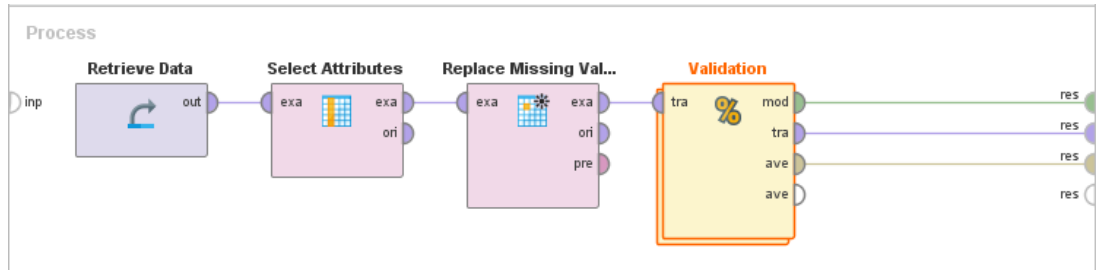


Figure 4-5: The main process of decision tree method in Rapid Miner tool for (MI) complication prediction

(Figure 4.5) includes multiple steps; the steps can be listed as follow:

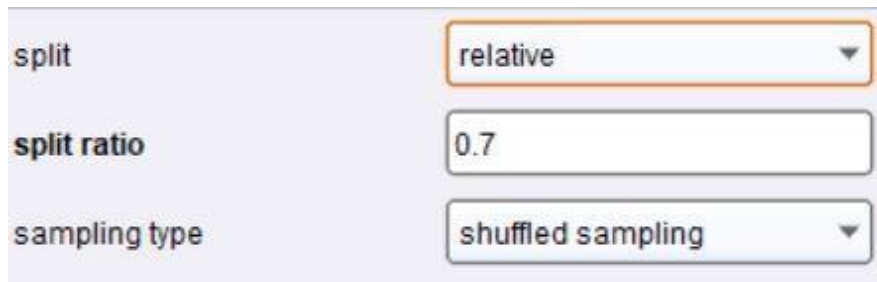
- Retrieve: this step retrieves the dataset from data repository after reading it from the excel file which consists of 785 instances belongs to patients who are infected with (MI).
- Replace Missing Values: replaces missing values of selected attributes by a specified replacement (Kalra & Aggarwal, 2017). In our experiment, we replace a missing value with 0 as a replacement value for the following attributes (“CHF”, “Stroke”, “ESRF”, “Blindness”, “Amputation”, “Age”, “EBP”, “DM”, “Score Smoking”, “LD”, “Score BMI”, “Inactivity”, “Family History”, “Result1”).
- Date to Numerical: at this step the operator changes the type of the selected date attribute to a numeric type (Kalra & Aggarwal, 2017). In this experiment, we change “Patient DOB” date attribute to represents as the year the patient was born.
- Discretize by User Specification: This operator discretizes the selected numerical attributes into user specified classes, the user can define the classes by specifying the upper limits of each class (Kalra & Aggarwal, 2017). We discretized the attribute “Patient DOB” (after using a date to numerical operator to represents the date values as the year the patients were born) into three classes.

- **Filter Examples:** This Operator selects which instances of a dataset will be kept according to the matching of the given condition defined by user, and which instances will be removed (Kalra & Aggarwal, 2017). In our experiment, we want to keep all instances with a value greater than 0 for “Result1” (blood glucose measurement) attribute.
- **Discretize by User Specification:** We discretized the attribute “Result1” (blood glucose measurement) into three classes (Danaei et al., 2011):
 - Low: With upper limit 70 mg/dl.
 - Normal: From 70 -130 mg/dl.
 - High: More than 200 mg/dl.
- **Select Attributes:** This Operator selects a subset of attributes of a dataset and removes the others (Kalra & Aggarwal, 2017). Because this process used for building a classifier to predict if the patient will be infecting with (MI), we choose the attributes “Patient Gender”, “Age”, “EBP”, “DM”, “Score Smoking”, “LD”, “Score BMI”, “Inactivity”, “Family History”, “Result1” and “MI”, and leave the attributes “CHF”, “Stroke”, “ESRF”, “Blindness” and “Amputation”, which refers to other complications.
- **Numerical to Binominal:** This operator changes the type of the selected numeric attributes to a binominal type. This operator also maps all values of these attributes to corresponding binominal values (Kalra & Aggarwal, 2017). In this process we choose the “MI” attribute to change its numeric type values to binominal type values, because it’s the predictable attribute in this process.
- **Set Role:** This operator is used to change the role of one or more attributes, all attributes in the dataset is a regular attribute by default. when we change an attribute to any other role, it's called a special attribute (Kalra & Aggarwal, 2017). In our process we changed the target role of the attribute “MI” to “label” as shown in [\(Figure 4.6\)](#), which means it’s a class label of our experiment.



Figure 4-6: Change “MI” attribute to a class label

- Split Validation: This operator randomly splits up the dataset into a training set and test set and evaluates the model. This operator performs a split validation in order to estimate the accuracy of the model (Kalra & Aggarwal, 2017). In our experiment, we choose a split ratio equal 0.7 as shown in [\(Figure 4.7\)](#), means that 70% of the dataset instances are used as a training set and the other 30% of instances are used as a test set.



split	relative
split ratio	0.7
sampling type	shuffled sampling

[Figure 4-7: Split ratio](#)

- Decision Tree: This Operator generates a decision tree model, which is used for classification.
- Apply Model: This operator applies a model on a dataset. While the decision tree operator used for generates the classification model based on the training set, the apply model get a prediction on the test set (Kalra & Aggarwal, 2017). [\(Figure 4.8\)](#).

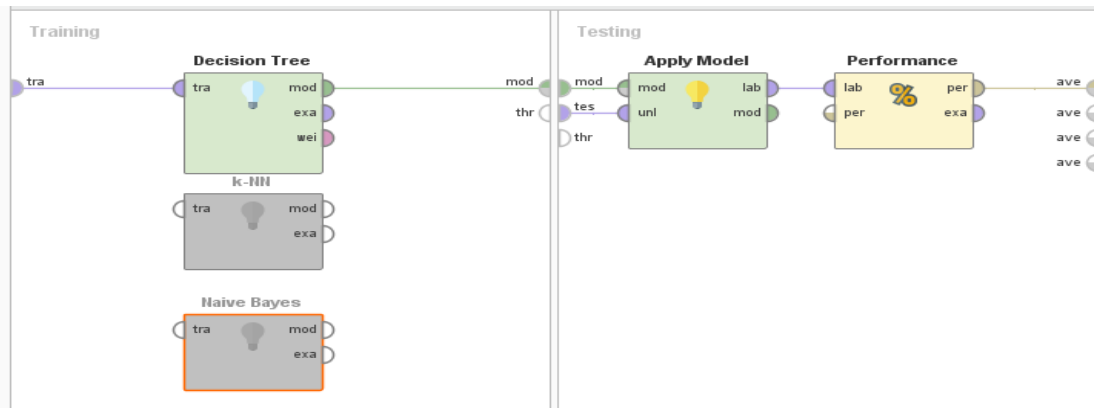


Figure 4-8: Decision Tree method

- Performance Classification: This operator is used for statistical performance evaluation of classification tasks (Kalra & Aggarwal, 2017). in our experiment, we measure performance accuracy for the classification model.

4.4.2 KNN

The following process (Figure 4.9) illustrated the main process of the KNN method. KNN method applied on the dataset to build a classifier for predicting if the patient will be infected with (MI), this method is implemented via Rapid Miner tool:

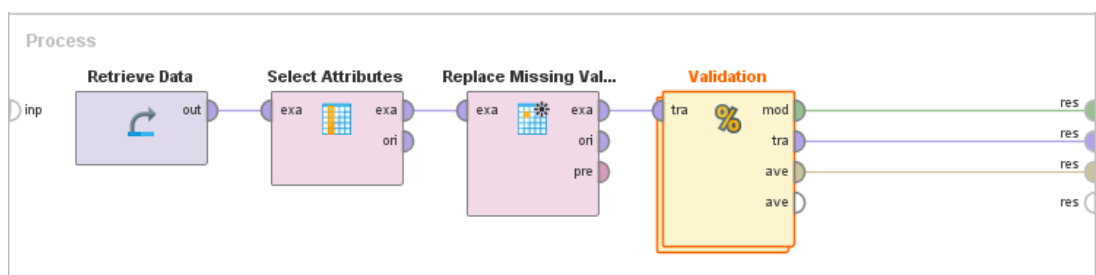


Figure 4-9: The main process of KNN in Rapid Miner tool for (MI) prediction

(Figure 4.9) includes multiple steps; the steps can be listed as follow:

- Retrieve: this step retrieves the dataset from data repository after reading it from the excel file which consists of 8001 instances belongs to patients who are infected with (MI).

- **Replace Missing Values:** replaces missing values of selected attributes by a specified replacement. In our experiment, we replace a missing value with 0 as a replacement value for the following attributes (“CHF”, “Stroke”, “ESRF”, “Blindness”, “Amputation”, “Age”, “EBP”, “DM”, “Score Smoking”, “LD”, “Score BMI”, “Inactivity”, “Family History”, “Result1”).
- **Date to Numerical:** at this step the operator changes the type of the selected date attribute to a numeric type (Kalra & Aggarwal, 2017). In this experiment, we change “Patient DOB” date attribute to represents as the year the patient was born.
- **Discretize by User Specification:** This operator discretizes the selected numerical attributes into user specified classes, the user can define the classes by specifying the upper limits of each class(Kalra & Aggarwal, 2017). We discretized the attribute “Patient DOB” (after using a date to numerical operator to represents the date values as the year the patients were born) into three classes.
- **Filter Examples:** This Operator selects which instances of a dataset will be kept according to the matching of the given condition defined by user, and which instances will be removed (Kalra & Aggarwal, 2017). In our experiment, we want to keep all instances with a value greater than 0 for “Result1” (blood glucose measurement) attribute.
- **Discretize by User Specification:** We discretized the attribute “Result1” (blood glucose measurement) into three classes:
- **Select Attributes:** This Operator selects a subset of attributes of a dataset and removes the others. Because this process used for building a classifier to predict if the patient will be infecting with (MI), we choose the attributes “Patient Gender”, “Age”, “EBP”, “DM”, “Score Smoking”, “LD”, “Score BMI”, “Inactivity”, “Family History”, “Result1” and “MI”, and leave the attributes “CHF”, “Stroke”, “ESRF”, “Blindness” and “Amputation”, which refers to other complications.
- **Numerical to Binominal:** This operator changes the type of the selected numeric attributes to a binominal type. This operator also maps all values of these attributes to corresponding binominal values (Kalra & Aggarwal, 2017). In this

process we choose the “MI” attribute to change its numeric type values to binominal type values, because it’s the predictable attribute in this process.

- Set Role: This operator is used to change the role of one or more attributes, all attributes in the dataset is a regular attribute by default. when we change an attribute to any other role, it's called a special attribute (Kalra & Aggarwal, 2017). In our process we changed the target role of the attribute “MI” to “label” as shown in [\(Figure 4.10\)](#), which means it’s a class label of our experiment.



Figure 4-10: Change “MI” attribute to a class label

- Split Validation: This operator randomly splits up the dataset into a training set and test set and evaluates the model. This operator performs a split validation in order to estimate the accuracy of the model (Kalra & Aggarwal, 2017). In our experiment, we choose a split ratio equal 0.7 as shown in [\(Figure 4.11\)](#), means that 70% of the dataset instances are used as a training set and the other 30% of instances are used as a test set.

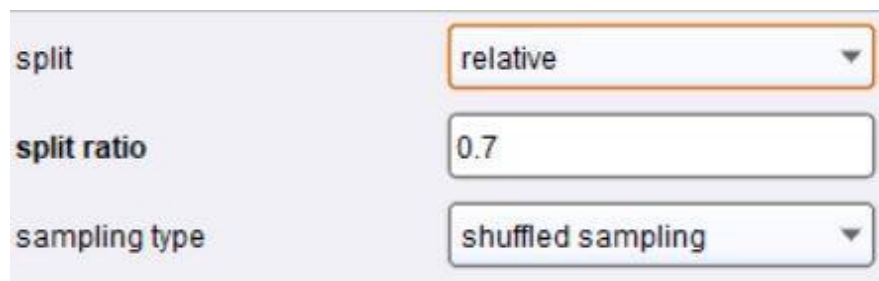


Figure 4-11: Split ratio

- KNN: This Operator generates a KNN model, which is used for classification.
- Apply Model: This operator applies a model on a dataset. While the KNN operator used for generates the classification model based on the training set, the apply model get a prediction on the test set (Kalra & Aggarwal, 2017). [\(Figure 4.12\)](#).

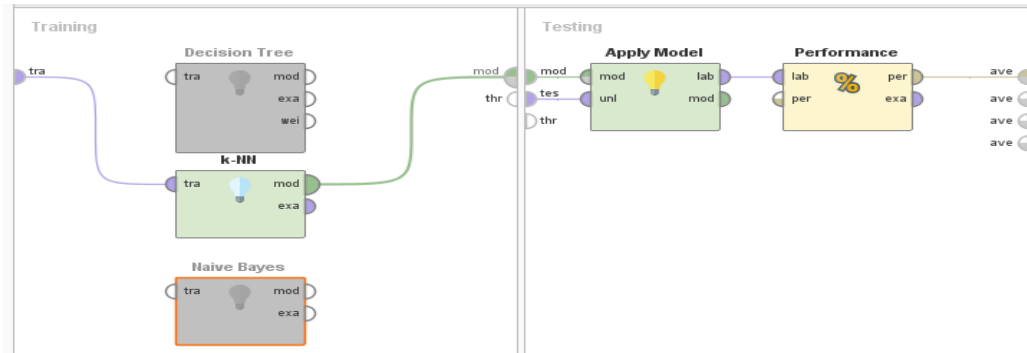


Figure 4-12: KNN method

- Performance Classification: This operator is used for statistical performance evaluation of classification tasks (Kalra & Aggarwal, 2017). in our experiment, we measure performance accuracy for the classification model.

4.4.3 Naïve Bayes

The following process (Figure 4.13) illustrated the main process of the Naïve Bayes method. NB method applied on the dataset to build a classifier for predicting if the patient will be infected with (MI), this method is implemented via Rapid Miner tool:

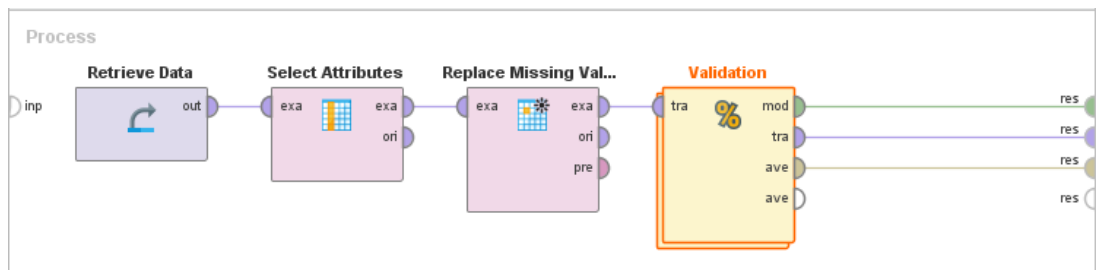


Figure 4-13: The main process of naïve bayesen method in Rapid Miner tool for Myocardial Infarction (MI)—prediction

(Figure 4.13) includes multiple steps; the steps can be listed as follow:

- Retrieve: this step retrieves the dataset from data repository after reading it from the excel file which consists of 785 instances belongs to patients who are infected with (MI).
- Replace Missing Values: replaces missing values of selected attributes by a specified replacement (Kalra & Aggarwal, 2017). In our experiment, we replace a missing value with 0 as a replacement value for the following attributes (“CHF”, “Stroke”, “ESRF”, “Blindness”, “Amputation”, “Age”, “EBP”, “DM”,

“Score Smoking”, “LD”, “BMI Score”, “Inactivity”, “Family History”, “Result1”).

- Date to Numerical: At this step the operator changes the type of the selected date attribute to a numeric type (Kalra & Aggarwal, 2017). In this experiment, we change “Patient DOB” date attribute to represents as the year the patient was born.
- Discretize by User Specification: This operator discretizes the selected numerical attributes into user specified classes, the user can define the classes by specifying the upper limits of each class. We discretized the attribute “Patient DOB” (after using a date to numerical operator to represents the date values as the year the patients were born) into three classes.
- Filter Examples: This Operator selects which instances of a dataset will be kept according to the matching of the given condition defined by user, and which instances will be removed. In our experiment, we want to keep all instances with a value greater than 0 for “Result1” (blood glucose measurement) attribute.
- Discretize by User Specification: We discretized the attribute “Result1” (blood glucose measurement) into three classes.
- Select Attributes: This Operator selects a subset of attributes of a dataset and removes the others (Kalra & Aggarwal, 2017). Because this process used for building a classifier to predict if the patient will be infecting with (MI), we choose the attributes “Patient Gender”, “Age”, “EBP”, “DM”, “Score Smoking”, “LD”, “Score BMI”, “Inactivity”, “Family History”, “Result1” and “MI”, and leave the attributes “CHF”, “Stroke”, “ESRF”, “Blindness” and “Amputation”, which refers to other complications.
- Numerical to Binominal: This operator changes the type of the selected numeric attributes to a binominal type. This operator also maps all values of these attributes to corresponding binominal values (Kalra & Aggarwal, 2017). In this process we choose the “MI” attribute to change its numeric type values to binominal type values, because it’s the predictable attribute in this process.
- Set Role: This operator is used to change the role of one or more attributes, all attributes in the dataset is a regular attribute by default. when we change an attribute to any other role, it's called a special attribute (Kalra & Aggarwal,

2017). In our process we changed the target role of the attribute “MI” to “label” as shown in (Figure 4.14), which means it’s a class label of our experiment.



Figure 4-14: Change “MI” attribute to a class label

- Split Validation: This operator randomly splits up the dataset into a training set and test set and evaluates the model. This operator performs a split validation in order to estimate the accuracy of the model (Kalra & Aggarwal, 2017). In our experiment, we choose a split ratio equal 0.7 as shown in (Figure 4.15), means that 70% of the dataset instances are used as a training set and the other 30% of instances are used as a test set.

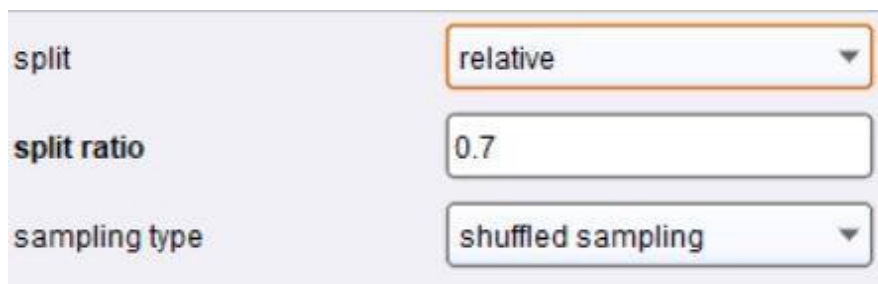


Figure 4-15: Split ratio

- Naïve Bayes: This Operator generates a model, which is used for classification.
- Apply Model: This operator applies a model on a dataset. While the Naïve Bayes operator used for generates the classification model based on the training set, the apply model get a prediction on the test set (Figure 4.16).

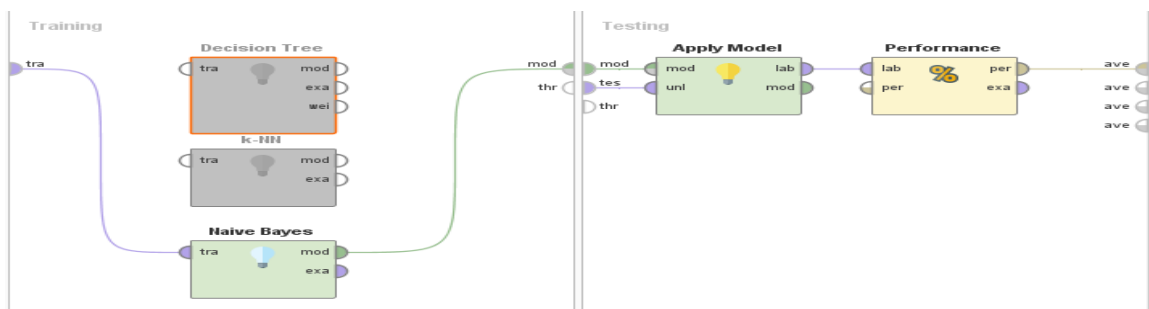


Figure 4-16: Naïve Baysen method

- Performance Classification: This operator is used for statistical performance evaluation of classification tasks (Kalra & Aggarwal, 2017). in our experiment, we measure performance accuracy for the classification model.

4.5 Summary

This chapter presents the experiments; the dataset attributes were discussed and the data preprocessing techniques was described, that used prior to the application of the data mining method. In addition, we presented the steps that have been performed to implement and evaluate the data mining system.

CHAPTER 5

Results and Discussion

CHAPTER 5

Results and Discussion

In this chapter, the results of applying data mining techniques are presented. Evaluating the used data mining techniques is based on the accuracy measure. This chapter is divided into 7 sections: the first section is the Decision tree results, the second section is showing the results of using KNN, the third section is about the results of Naive Bayesian, the fourth section is about result and conclusion, the fifth section about discussion, the sixth section talk about limitations of the study, the last section is the summary of the previous sections.

5.1 Decision Tree:

In this section, the decision tree model results for each MI rule. After applying the decision tree depending on splitting with the highest importance, the Rapid Miner studio generate the tree. The extracted rules from the experiment dataset after applying decision tree classification method on 8001 instances, which illustrate the factors leading to the probability of (MI) occurrence of are illustrated in [\(figure 5.1\)](#) and [\(table 5.1.\)](#)

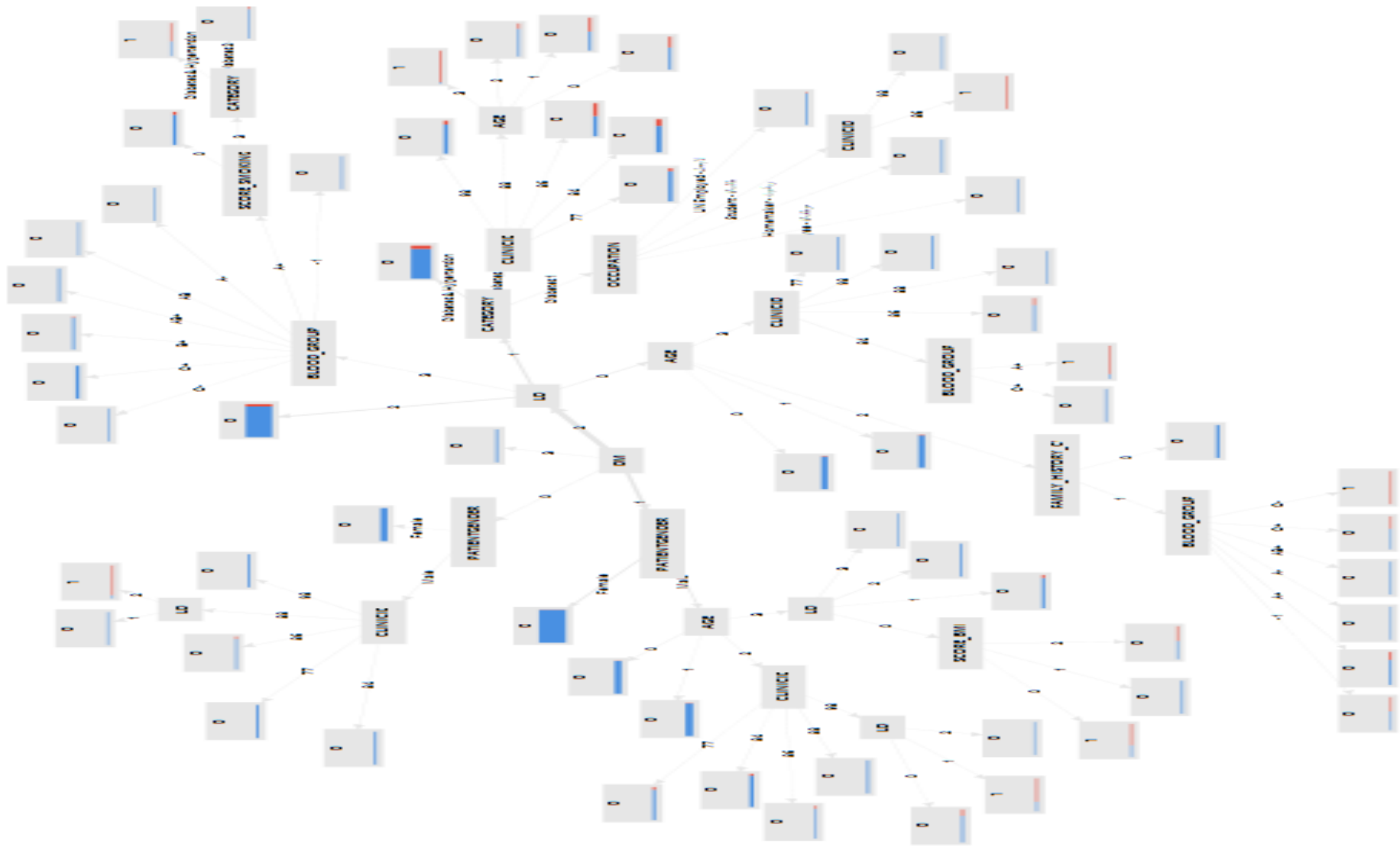


Figure 5-1: Myocardial Infarction (MI) decision tree output result

Table 5-1:Decision Tree Rules Table

Rule Number	Description
Rule 1	If ((DM = 0) And (PATIENTGENDER = male) And (clinic ID = 88) And (LD =2)) then the probability of Myocardial Infarction (MI) infection = true
Rule 2	If ((DM = 1) And (PATIENTGENDER = male) And (Age= 2) And (clinic ID = 98) AND (LD = 1)) then the probability of Myocardial Infarction (MI) infection = true
Rule 3	If ((DM = 1) And (PATIENTGENDER = male) And (Age= 3) And (LD =0) and (SCORE_BM=0)) then the probability of Myocardial Infarction (MI) infection = true
Rule 4	If ((DM = 2) And (LD =0) And (Age=2) And(FAMILY_HISTORY_CV =1) And (BLOOD_GROUP= O-)) then the probability of Myocardial Infarction (MI) infection = true
Rule 5	If ((DM = 2) And (LD =0) And (Age=3) And (clinic ID = 84) And (BLOOD_GROUP= A+)) then the probability of Myocardial Infarction (MI) infection = true
Rule 6	If ((DM = 2) And (LD =1) And (Category = Diabetes 1) And(OCCUPATION= Student) And (CLINICID= 86)) then the probability of Myocardial Infarction (MI) infection = true
Rule 7	If ((DM = 2) And (LD =1) And (Category = Diabetes 2) And (CLINICID= 88) And (Age =3)) then the probability of Myocardial Infarction (MI) infection = true
Rule 8	If ((DM = 2) And (LD =3) And (BLOOD_GROUP= A+) And (SCORE_SMOKING=3) And (Category = Diabetes &hypertension)) then the probability of Myocardial Infarction (MI) infection = true

The extracted rules above are showing that:

- Rule 1 means that the possibility of MI occurrence is true if the person has no Diabetes Mellitus, the gender is male, his Lipids disorders is Cholesterol 200 - 249 mg/dl and from Alsaftwai clinic.
- Rule 2 means that the possibility of MI occurrence is true if the person has Diabetes Mellitus Controlled with medications, gender is male, age is between 56 – 64 years and Cholesterol 160 - 199 mg/dl.
- Rule 3 means that the possibility of MI occurrence is true if the person has Diabetes Mellitus Controlled with medications, gender is male, older than 65 years old and $BMI \leq 29$.
- Rule 4 means that the possibility of MI occurrence is true if the person has Diabetes Mellitus Uncontrolled with medications, Cholesterol < 160 mg/dl, patient age between 56 – 64 years, family history is positive and Patient Blood Group is O-.
- Rule 5 means that the possibility of MI occurrence is true if the person has Diabetes Mellitus Uncontrolled with medications, Cholesterol < 160 mg/dl, age is over 65 years old, from Khan Younes and his blood group is A+.
- Rule 6 means that the possibility of MI occurrence is true if the person has Diabetes Mellitus Uncontrolled with medications, Cholesterol 160 - 199 mg/dl, the category is diabetes 1, student and from Alshaboura – Rafah.
- Rule 7 means that the possibility of MI occurrence is true if the person has Diabetes Mellitus Uncontrolled with medications, Cholesterol 160 - 199 mg/dl, the category is diabetes 2, from Alsaftawi and over 65 years old.
- Rule 8 means that the possibility of MI occurrence is true if the person has Diabetes Mellitus Uncontrolled with medications, Cholesterol ≥ 250 mg/dl, blood group is A+, heavy smoker and has both Diabetes & hypertension.

According to the 8 rules, the Diabetes Mellitus Uncontrolled with medications was shared in 5 rules among all, while Diabetes Mellitus Controlled with medications in 2 rules and only one rule the Diabetes Mellitus Controlled with lifestyle modification. the Age over 65 years old was also shared in 3 roles among all, while age between 56 – 64 years was in 2 rules. Blood Group appears A+ appears in 2 rules, while blood group O- appears in one rule only, the other 5 rules shows nothing related to blood groups. The category appears also in 3 rules; 1 Diabetes 1, 1 Diabetes 2, 1 Diabetes

&hypertension and the other 5 rules shows nothing related to category. Only one category out of 8 categories showed relation with heavy smoking.

```

Tree
DM = 0
| PATIENTGENDER = Female: 0 {0=310, 1=0}
| PATIENTGENDER = Male
| | CLINICID = 77: 0 {0=71, 1=0}
| | CLINICID = 84: 0 {0=52, 1=0}
| | CLINICID = 86: 0 {0=22, 1=2}
| | CLINICID = 88
| | | LD = 1: 0 {0=11, 1=0}
| | | LD = 2: 1 {0=1, 1=9}
| | CLINICID = 98: 0 {0=57, 1=0}
DM = 1
| PATIENTGENDER = Female: 0 {0=1409, 1=9}
| PATIENTGENDER = Male
| | AGE = 0: 0 {0=314, 1=2}
| | AGE = 1: 0 {0=351, 1=5}
| | AGE = 2
| | | CLINICID = 77: 0 {0=79, 1=7}
| | | CLINICID = 84: 0 {0=142, 1=8}
| | | CLINICID = 86: 0 {0=18, 1=2}
| | | CLINICID = 88: 0 {0=50, 1=0}
| | | CLINICID = 98
| | | | LD = 0: 0 {0=38, 1=9}
| | | | LD = 1: 1 {0=5, 1=12}
| | | | LD = 2: 0 {0=4, 1=0}
| | AGE = 3
| | | LD = 0
| | | | SCORE_BMI = 0: 1 {0=6, 1=11}
| | | | SCORE_BMI = 1: 0 {0=30, 1=0}
| | | | SCORE_BMI = 2: 0 {0=10, 1=8}
| | | LD = 1: 0 {0=87, 1=8}
| | | LD = 2: 0 {0=42, 1=0}
| | | LD = 3: 0 {0=6, 1=0}

```

Figure 5-2: Part of Decision Tree predictor

Figure 5.2 illustrating a part the Decision Tree predictor, the figure showed the results of each rule; if DM is 0, gender is male and the clinic id is 22 there were 2 patients have MI. 9 patients have MI where with 0 DM, males, clinic id is 88 and LD is 2, to the end of the figure.

The accuracy of using the decision tree algorithm was 92.54% with 48.48% precision and 8.99% recall as shown in following [Figure 5.3](#).

accuracy: 92.54%

	true 0	true 1	class precision
pred. 0	2205	162	93.16%
pred. 1	17	16	48.48%
class recall	99.23%	8.99%	

Figure 5-3: Decision Tree accuracy

5.2 K-Nearest Neighbor

In this section. KNN model is tested, and the KNN model is applied 5 times with editing the K value in each time. The following table (Table 5-2) listed the Accuracy for each attempt when applying a different K Value.

Table 5-2: KNN Accuracy with different K Value

K	Accuracy
1	75.96%
2	88.83%
3	88.42%
4	97%
5	96.00%

According to the table above, the highest accuracy recorded when the K value was 4 with 86.38% followed by 86.00% when the K value was 5, then 82.42% recorded with 3 K value, after that 82.83% when the K value was 2 and finally with the least recorded accuracy by 75.96% with 1 K value.

Figure 5.4 illustrating the KNN Accuracy with K value 4 which record the highest accuracy. The figure shows that class precision was 98.31% when the prediction is 0 and 89.76% when the prediction is 1, while the class recall was 99.28% when for 0 cases and 78.65% when cases was 1.

accuracy: 97.75%

	true 0	true 1	class precision
pred. 0	2206	38	98.31%
pred. 1	16	140	89.74%
class recall	99.28%	78.65%	

Figure 5-4: KNN Accuracy with K value 4

5.3 Naïve Bayes:

Applying the naïve Bayes techniques on the dataset shows different results from the previous techniques. Figure 5.5 shows the accuracy using naïve Bayes was 93%,

while the class precision recorded when the prediction is 0 was 94.42% and 55.58% when the prediction is 1, the class recall was 98.24% for the MI 0 and 27.53% when the MI is 1.

accuracy: 93.00%

	true 0	true 1	class precision
pred. 0	2183	129	94.42%
pred. 1	39	49	55.68%
class recall	98.24%	27.53%	

Figure 5-5: naive bayes accuracy

According to the follow figure the density distribution in table as see in the [following Table 5-3:](#)

Attribute	Parameter	MI=0	MI=1
CLINIC ID	VLAUIE=86	0.095	0.143
CLINIC ID	VLAUIE=88	0.142	0.175
CATEGORY	VALUE=HYPERTENTION	0.002	0.000
Patient Gender	Value =F mail	0.507	0.452
Patient Gender	Value =mail	0.493	0.548
Occupation	Value = employee	0.131	0.180
DM	Value=2	5.78	0.734
Family history	Value=2	0.647	0.764
Score BMI	Value=0	0.273	0.304

Table 5-3: Distribution for MI in fields with more satisfactory density.

5.4 Results conclusion

The accuracy of the three algorithms is shown in [Table 5.3](#), the highest accuracy recorded with KNN algorithm with 95.46% followed with Naïve Bayes with 93%, where the Decision Tree recorded the 92.54% in the third place.

Also, the KNN algorithm recorded the highest retrieved instances by the classifier with 78.99% as the precision for KNN use, followed with Naïve Bayes with 55.65% and Decision tree place as the last precision with 48.48%, finally the fraction of relevant instances that are retrieved by the classifier for KNN also was the highest among the three algorithms with 52.81% recall followed by Naïve Bayes with 27.53% where the Decision Tree recorded 8.99%. Based on findings the KNN scored the best among the used algorithms, which means it is the best algorithm to use for our predictive model.

Table 5-4: Accuracy List

No.	Model	Accuracy	Precision	Recall
1	Decision Tree	92%	48.48%	8.99%
2	K-Nearest Neighbor _(K=4)	97%	78.99%	52.81%
3	Naïve Bayes	93%	55.65%	27.53%

Graphical representation of accuracy for each algorithm in both cases is given below [\(Figure 5-6\)](#):

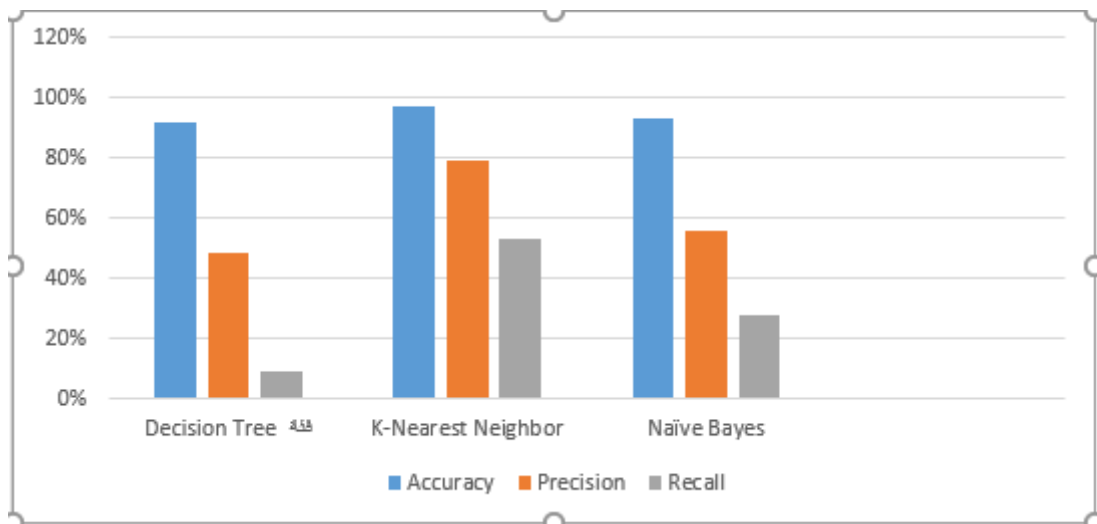


Figure 5-6: Graphical representation of the accuracy, Precision and Recall for each classifier

5.5 Discussion

This study aimed to assess the predictive ability of MI using data mining techniques among patients with chronic diseases. Data have been processed by using data mining algorithms through three techniques decision trees, KNN, and Naïve Bayes algorithm for prediction to extract the probability of a heart attack from combined risk factors. The accuracy of the three algorithms is the highest accuracy recorded with KNN algorithm with 95.46% followed with Naïve Bayes with 93%, where the Decision Tree recorded the 92.54% in the third place.

The study results came in line with what have been concluded in Bangladesh by Nag, Procheta, et al which predicted the risk of heart attack for a person complaining of chest pain or equivalent symptoms (Nag et al., 2017). In spite of the difference of data mining methods that have been used we reached to the same conclusion. Since they used c4.5 decision tree algorithm, which its accuracy improved by random forest, we used decision tree, KNN, and Naïve Bayes algorithms.

The situation was not much different in the study performed by (Safdari et al., 2013) except in the accuracy of decision tree algorithms result. They used data mining algorithms with multi model (C5, C&RT, QUEST, CHAD) with accuracy score of (85.7%, 79.1%, 72.28%, 93.4%) respectively. While, in our study we used data mining algorithms by the decision tree model which gained accuracy of 92.54%. In addition, we applied KNN and Naïve Bayes algorithms as a sensitivity tests which gave similar accuracy of (95.46%, 93%) respectively. Therefore, we conclude that KNN was an applicable algorithm for clinical purposes with the highest precision.

As well, the study conducted by Priyanka et al which used the Naïve Bayes for heart diseases prediction and scored 82.3% accuracy (Priyanka, N., & RaviKumar, P. 2017), while the accuracy of algorithm for predicting MI scored 98.8% with (Arif, 2010). Still, we recorded 93% accuracy for Naïve Bayes in our study.

However, data mining is a robust technique for extracting predictive information from the extensive databases. In the study conducted by Hasan et al by using info gain feature selection technique and removing unnecessary features. Different classification techniques such that KNN, Decision Tree (ID3), Gaussian Naïve Bayes, Logistic Regression and Random Forest are used on heart disease dataset for better prediction. They considered different performance measurement factors such as accuracy, ROC curve, precision, recall, sensitivity, specificity, and F1-score are to determine the

performance of the classification techniques. They concluded that Logistic Regression performed better (with a classification accuracy of 92.76%) among all approaches (Hasan et al., 2018).

Therefore, feature selection technique and removing unnecessary features seems to be a high necessity consideration for improve data mining techniques based on medical considerations.

5.6 Limitations of the study

In spite of the good results we have reached in this study, several limitations float as follow:

- Feature selection technique was done based on results and better accuracy.
- The dataset was collected from UNRWA primary health centers in the period 2015.
- Data were extracted only from five UNRWA primary health centers in Gaza Strip (Khan Younis, Al Nosirat, Al Remal, Saftawy, Rafah) out of twenty-two.
- Our work focuses only on patients diagnosed with chronic diseases.
- Not all Data mining techniques were used.

5.7 Summary

This chapter described experiments results of predicting MI based on many factors such as clinic ID, history of diabetes & hypertension, gender, occupation, blood group, family history, age, smoking score, lipids disorders and BMI score. Three data mining techniques were used, as following: Decision tree, K Nearest Neighbor, and Naïve Bayesian. Each technique give different results based on data nature and attempts number had done on the data set.

The accuracy for the three used techniques (Decision Tree, Naïve Bayesian, and K Nearest Neighbor) were relatively close. Clearly, the highest accuracy was for KNN followed by Naïve Bayes then Decision tree.

CHAPTER 6

Conclusion and Recommendations

CHAPTER 6

Conclusion and Recommendations

This chapter consist of four sections, section 6.1 presents research summary, section 6.2 presents the useful conclusions about the research experiments, section 6.3 presents useful recommendation which extracted from this thesis about Prediction MI among chronic diseases, finally section 6.4 presents possible directions of future works.

6.1 Thesis summary

In this research, a data mining techniques has been used for Predicting (MI) among chronic diseases patients. The research aims to predict the factors causes (MI) among chronic diseases patients. The patients' data were collected from UNRWA Clinics in Gaza Strip.

Our methodology in this thesis consisted of several steps: first, understand and prepare the data set collected from UNRWA clinics, in the second step we implemented data mining methods to build a classifier for extracting useful rules from the data set, and the third step, evaluated the data mining system, in the fourth step, the extracted rules from data mining system to predict the complications for MI patients based on results from the data mining techniques. The fifth step, we discussed the results.

The results of this study indicates that the accuracy of the data mining techniques. The accuracy was 92.54% for decision tree, 95.46% for KNN and 93% for Naïve Bayes, the accuracy gives us accurate results in predicting MI occurrence for any patient.

6.2 Conclusion

The following conclusions can be drawn about the MI prediction among the chronic diseases patients. First, the clinics and the health departments should focus on using data mining techniques for MI prediction. The use of data mining techniques could offer a huge step for predicting the MI occurrence for the chronic diseases patients which prevent a huge complication and life loses. In other words, the patient may have chronic diseases and this chronic disease may play a role factor for MI, based on that, patient will have MI with serious complications may threat their lives.

The results of this research enable the prediction of MI among chronic diseases patients that they may avoid the risk on their life based on the 5 UNRWA clinics. The

data have been processed to predict the MI using 3 data mining techniques; decision tree, KNN and naïve Bayes.

According to the results in chapter 5. The chronic diseases such as DM, LD, Diabetes, hypertension and other factors such as the blood group and age are playing main role in predicting the MI occurrence. The accuracy in the three data mining techniques is high and close enough.

6.3 Recommendations

Based on the findings and the conclusions of the study, the following will be recommendations to be considered:

1. Feature selection technique must have done based on medical considerations.
2. Patients should care about their health situation and their lifestyle, eating unhealthy foods for preventing LD and care about BMI. In addition, they should try to reduce the factors that lead to MI occurrence and care about their DM status.
3. Expanding the data collection circle to include both chronic and non-chronic diseases.
4. Cooperation between health institutions and agencies to carry out awareness programs aimed at raising citizens' awareness of the relation between MI and chronic diseases.
5. Collaborate among specialists to improve system accuracy by working to find other factors that may predict the MI.
6. Activating the work of information technology in health institutions, especially data mining, because it is important to extract hidden knowledge in all fields.
7. UNRWA and other health institutions are advised to take the results of this study in consideration and expand the study to predict MI among chronic diseases patients which going to help in saving lives.
8. We also recommend expanding the data sets and have more specified data for MI and other chronic diseases.

6.4 Future Works

Since this study had only focused on predicting (MI) among chronic diseases patients using data mining to a collected data from UNRWA clinics in Gaza Strip, it is recommended that future works be carried out on the following possible directions:

1. Combining other techniques with data mining methods Particularly Logistic regression to extract more accurate rules.
2. To extract new knowledge about MI and chronic diseases from all other health institutions in Gaza Strip and West Bank.
3. To use other data mining methods to extract new knowledge about MI such as association rules.
4. To expand the data sets from all health institution and more details to measure the accuracy from bigger data sets.
5. Working on applications for patients, which giving them tips and hints to care about their heath and to know their health status.
6. Working on models to predict other chronic diseases.

REFERENCES

- Abed, Y., & Jamee, A. (2015). Characteristics and risk factors attributed to coronary artery disease in women attended health services in Gaza-Palestine observational study. *World Journal of Cardiovascular Diseases*, 5(01), 9.
- Anderson, T. J., Grégoire, J., Hegele, R. A., Couture, P., Mancini, G. J., McPherson, R., . . . Grover, S. (2013). 2012 update of the Canadian Cardiovascular Society guidelines for the diagnosis and treatment of dyslipidemia for the prevention of cardiovascular disease in the adult. *Canadian Journal of Cardiology*, 29(2), 151-167.
- Apple, F. S., Wu, A. H., Mair, J., Ravkilde, J., Panteghini, M., Tate, J., . . . Danne, O. (2005). Future biomarkers for detection of ischemia and risk stratification in acute coronary syndrome. *Clinical Chemistry*, 51(5), 810-824.
- Bellamy, L., Casas, J.-P., Hingorani, A. D., & Williams, D. J. (2007). Pre-eclampsia and risk of cardiovascular disease and cancer in later life: systematic review and meta-analysis. *Bmj*, 335(7627), 974.
- Borro, L. C., Oliveira, S. R., Yamagishi, M. E., Mancini, A. L., Jardine, J. G., Mazoni, I., . . . Neshich, G. (2006). Predicting enzyme class from protein structure using Bayesian classification. *Embrapa Informática Agropecuária-Artigo em periódico indexado (ALICE)*.
- Braunwald, E., Fauci, A., Kasper, D., Hauser, S., Longo, D., & Jameson, J. (2001). *Harrisons Principles of Internal Medicine* McGraw-Hill. New York, 1377-1385.
- Brown, D., Bayley, I., Harrison, R., & Martin, C. (2013). *Developing a mobile case-based reasoning application to assist type 1 diabetes management*. Paper presented at the 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013).
- Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 1165-1188.
- Danaei, G., Finucane, M. M., Lu, Y., Singh, G. M., Cowan, M. J., Paciorek, C. J., . . . Stevens, G. A. (2011). National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants. *The lancet*, 378(9785), 31-40.
- Denmark, K. T., Bax, J. J., Morrow, D. A., Task, A., Kristian, C., Denmark, T., . . . Germany, H. A. K. (2019). Fourth universal definition of myocardial infarction (2018).
- Detection, N. C. E. P. E. P. o., & Adults, T. o. H. B. C. i. (2002). *Third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III)*: National Cholesterol Education Program, National Heart, Lung, and Blood . . .
- Everson-Rose, S. A., Lewis, T. T., Karavolos, K., Dugan, S. A., Wesley, D., & Powell, L. H. (2009). Depressive symptoms and increased visceral fat in middle-aged women. *Psychosomatic medicine*, 71(4), 410.
- Fadl Elmula, F. E. M., Hoffmann, P., Larstorp, A. C., Fossum, E., Brekke, M., Kjeldsen, S. E., . . . Rostrup, M. (2014). Adjusted drug treatment is superior to renal sympathetic denervation in patients with true treatment-resistant hypertension. *Hypertension*, 63(5), 991-999.

- García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1), 9.
- Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Blaha, M. J., . . . Franco, S. (2014). Executive summary: heart disease and stroke statistics—2014 update: a report from the American Heart Association. *Circulation*, 129(3), 399-410.
- Gorade, S., Deo, A., & Purohit, P. (2017). A study of some data mining classification techniques. *International Research J. of Engineering and Technology (IRJET)*, 4.
- Grundy, S. M., Cleeman, J. I., Merz, C. N. B., Brewer, H. B., Clark, L. T., Hunnigake, D. B., . . . Program, C. C. o. t. N. C. E. (2004). Implications of recent clinical trials for the national cholesterol education program adult treatment panel III guidelines. *Journal of the American College of Cardiology*, 44(3), 720-732.
- Han, J., Kamber, M., & Pei, J. J. M. K. P. (2012). Data mining: concepts and techniques, Waltham, MA. 10, 978-971.
- Hasan, S., Mamun, M., Uddin, M., & Hossain, M. (2018). *Comparative Analysis of Classification Approaches for Heart Disease Prediction*. Paper presented at the 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2).
- Huxley, R., Barzi, F., & Woodward, M. (2006). Excess risk of fatal coronary heart disease associated with diabetes in men and women: meta-analysis of 37 prospective cohort studies. *BMJ*, 332(7533), 73-78.
- Jamee, A., & Abed, Y. (2014). Coronary Artery Disease in Overweight and Obese Women in Gaza-Palestine: An Observational Study. *American Journal of Cardiovascular Disease Research*, 2(2), 23-26.
- Jennings, R., & Ganote, C. E. (1974). Structural changes in myocardium during acute ischemia. *Circulation Research*, 35(3_supplement), III-156-III-172.
- Kalra, V., & Aggarwal, R. (2017). *Importance of Text Data Preprocessing & Implementation in RapidMiner*. Paper presented at the Proceedings of the First International Conference on Information Technology and Knowledge Management—New Dehli, India.
- Kannel, W. B., & Vasan, R. S. (2009). Triglycerides as vascular risk factors: new epidemiologic insights for current opinion in cardiology. *Current opinion in cardiology*, 24(4), 345.
- Kessous, R., Shoham-Vardi, I., Pariente, G., Sherf, M., & Sheiner, E. (2013). An association between gestational diabetes mellitus and long-term maternal cardiovascular morbidity. *Heart*, 99(15), 1118-1121.
- Kolçe, E., & Frasher, N. (2012). A literature review of data mining techniques used in healthcare databases. *ICT innovations*.
- Kones, R. (2011). Primary prevention of coronary heart disease: integration of new data, evolving views, revised goals, and role of rosuvastatin in management. A comprehensive survey. *Drug design, development and therapy*, 5, 325.
- Lior, R. (2014). *Data mining with decision trees: theory and applications* (Vol. 81): World scientific.
- Longo, D. L., Fauci, A. S., Kasper, D. L., Hauser, S. L., Jameson, J. L., & Loscalzo, J. (2012). *Harrison's principles of internal medicine* (Vol. 2012): Mcgraw-hill New York.

- Maimon, O., & Rokach, L. (2009). Introduction to knowledge discovery and data mining. In *Data mining and knowledge discovery handbook* (pp. 1-15): Springer.
- Marín-García, J., & Goldenthal, M. J. (2008). Mitochondrial centrality in heart failure. *Heart failure reviews*, 13(2), 137-150.
- Mason, J. C., & Libby, P. (2015). Cardiovascular disease in patients with chronic inflammation: mechanisms underlying premature cardiovascular events in rheumatologic conditions. *European heart journal*, 36(8), 482-489.
- McGowan, E. C., & Keet, C. A. (2013). Prevalence of self-reported food allergy in the National Health and Nutrition Examination Survey (NHANES) 2007-2010. *Journal of Allergy and Clinical Immunology*, 132(5), 1216-1219. e1215.
- Mosca, L., Barrett-Connor, E., & Kass Wenger, N. (2011). Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes. *Circulation*, 124(19), 2145-2154.
- Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., . . . Fullerton, H. J. (2015). Aha statistical update. *Circulation*, 132, 000-000.
- Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., . . . Howard, V. J. (2015). Executive summary: heart disease and stroke statistics—2015 update: a report from the American Heart Association. *Circulation*, 131(4), 434-441.
- Nag, P., Mondal, S., Ahmed, F., More, A., & Raihan, M. (2017). *A simple acute myocardial infarction (Heart Attack) prediction system using clinical data and data mining techniques*. Paper presented at the 2017 20th International Conference of Computer and Information Technology (ICCIIT).
- Oded, & Maimon. (2010). *Data mining and knowledge discovery Handbook*. 875.
- Oded, M., & Lior, R. (2010). *Data mining and knowledge discovery Handbook. chapter 45: Data Mining for Imbalanced Datasets: An Overview.*, 875.
- PAIC. (2017). Health Annual Report, Palestine 2016. *Ministry of Health, Palestine*.
- Payne, R. A. (2012). Cardiovascular risk. *British journal of clinical pharmacology*, 74(3), 396-410.
- Pezzè, M., Young, M., Guan, K. C., Heng, D., Yong, T. T., Johnston, S. F., . . . King, W. J. (2012-2013). RECOMMENDED TEXTBOOKS AND REFERENCES FOR EEE FOURTH YEAR ACADEMIC YEAR 2012-2013 SEMESTER.
- Phyu, T. N. (2009). *Survey of classification techniques in data mining*. Paper presented at the Proceedings of the International MultiConference of Engineers and Computer Scientists.
- Ramachandran, A., Liu, Y., Asghar, W., & Iqbal, S. M. (2009). Characterization of DNA-nanopore interactions by molecular dynamics. *Am J Biomed Sci*, 1(4), 344-351.
- Safdari, R., Saedi, M. G., Arji, G., Gharooni, M., Soraki, M., & Nasiri, M. (2013). A model for predicting myocardial infarction using data mining techniques. *Frontiers in Health Informatics*, 2(4), 1-6.
- Sankaranarayanan, S., & Perumal, T. P. (2014). *A predictive approach for diabetes mellitus disease through data mining technologies*. Paper presented at the 2014 World Congress on Computing and Communication Technologies.
- Schnabel, R. B., Sullivan, L. M., Levy, D., Pencina, M. J., Massaro, J. M., D'Agostino Sr, R. B., . . . Tadros, T. M. (2009). Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *The Lancet*, 373(9665), 739-745.

- Schoenborn, C. A., & Stommel, M. (2011). Adherence to the 2008 adult physical activity guidelines and mortality risk. *American journal of preventive medicine*, 40(5), 514-521.
- Shahwan, A. J., Abed, Y., Desormais, I., Magne, J., Preux, P. M., Aboyans, V., & Lacroix, P. (2019). Epidemiology of coronary artery disease and stroke and associated risk factors in Gaza community–Palestine. *PloS One*, 14(1).
- Silveira, E. A. d., Vieira, L. L., Jardim, T. V., & Souza, J. D. d. (2016). Obesity and its association with food consumption, diabetes mellitus, and acute myocardial infarction in the elderly. *Arquivos brasileiros de cardiologia*, 107(6), 509-517.
- Singh, D. K., & Swaroop, V. (2013). Data security and privacy in data mining: research issues & preparation. *International Journal of Computer Trends and Technology*, 4(2), 194-200.
- Sun, Y., & Han, J. (2012). Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2), 1-159.
- Thygesen, K., Alpert, J. S., & Jaffe, A. S. (2013). Erratum: Third universal definition of myocardial infarction (Journal of the American College of Cardiology (2012) 60 (158-98). *Journal of the American College of Cardiology*, 61(5), 598.
- Thygesen, K., Alpert, J. S., Jaffe, A. S., Chaitman, B. R., Bax, J. J., Morrow, D. A., . . . Van de Werf, F. (2019). Fourth universal definition of myocardial infarction (2018). *European heart journal*, 40(3), 237-269.
- Thygesen, K., Alpert, J. S., White, H. D., TASK FORCE MEMBERS: Chairpersons: Kristian Thygesen , J. S. A., Harvey D. White *, Biomarker Group: Allan S. Jaffe, C., Fred S. Apple , Marcello Galvani , Hugo A. Katus , L. Kristin Newby , Jan Ravkilde, ECG Group: Bernard Chaitman, C.-o., Peter M. Clemmensen , Mikael Dellborg , Hanoch Hod , Pekka Porela, . . . Global Perspective Group: Philip A. Poole-Wilson, C., Enrique P. Gurfinkel , José-Luis Lopez-Sendon , Prem Pais , Shanti Mendis , Jun-Ren Zhu. (2007). Universal definition of myocardial infarction. *Circulation*, 116(22), 2634-2653.
- Valensi, P., Lorgis, L., & Cottin, Y. (2011). Prevalence, incidence, predictive factors and prognosis of silent myocardial infarction: a review of the literature. *Archives of Cardiovascular Diseases*, 104(3), 178-188.
- WHO. (2002). *The world health report 2002: reducing risks, promoting healthy life*: World Health Organization.
- Wolk, R., Berger, P., Lennon, R. J., Brilakis, E. S., & Somers, V. K. (2003). Body mass index: a risk factor for unstable angina and myocardial infarction in patients with angiographically confirmed coronary artery disease. *Circulation*, 108(18), 2206-2211.
- Wu, F. (2012). *Discussion on experimental teaching of data warehouse & data mining course for undergraduate education*. Paper presented at the 2012 7th International Conference on Computer Science & Education (ICCSE).
- Xing, Y., Wang, J., & Zhao, Z. (2007). *Combination data mining methods with new medical data to predicting outcome of coronary heart disease*. Paper presented at the 2007 International Conference on Convergence Information Technology (ICCIT 2007).
- Yeh, D.-Y., Cheng, C.-H., & Chen, Y.-W. (2011). A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications*, 38(7), 8970-8977.
- ZOLER, M. L. (2006). Five Types of MI Will Make Up New Definition. *Arthritis Rheum*, 54, 2688-2696.

