MEDICAL DIAGNOSIS USING DATA MINING TECHNIQUES

NAME	: SHAIFUL NIZAM B. ZAMRI
MATRIKS NUMBER	: WEK 990366
SUPERVISOR	: ASSOC. PROF. DR. N. SELVANATHAN
MODERATOR	: MR. WOO CHAW SENG

This project is submitted to the Faculty of Computer Science and Information Technology, University of Malaya

in partial fulfillment of the requirement for the Bachelor of Computer Science (Honors)

session 2002/2003.

Abstract

Medical Diagnosis using Data Mining Techniques is a system in specifically facilitates the data mining techniques in predicting the appropriate drug prescription of certain condition of patients. The report will firstly cover the inductors of data mining and the overview of the overall system and it also reviewed about the data mining paradigm.

Secondly, this report will review the literature part which started with basic knowledge of data mining and knowing what the basic information about data mining. At the end of this part will reviewed and studied about the data mining algorithm that gathered from the research made earlier.

Then followed by the methodology part as the scenario of the project and the system analysis, which will review the needed requirements and tools used to implement the system. The system design part will explains the flow and functionality of the system structure.

Then, the system implementation part explains the selected tools chosen for implementing the system which using the Clementine Data Mining Solution.

System testing then was done for checking and detecting error or malfunction of the system.

The final part is generally the overall system evaluation and the conclusion of the whole project which explains the strength and the limitation of the system.

Acknowledgement

Here, I would like to express my sincere appreciation to my supervisor, Assoc. Prof. Dr. Selvanathan Narainasamy. He has given the advice and full guidance during my research of this project.

Secondly, I am also grateful and thankful to my moderator Mr. Woo Chaw Seng for his useful advice, guidance and comment for ensuring the project will meet the requirements needed.

I am also very grateful to my group mates Nik Anuar Fathi and Mohd. Faizal, withal the co-operation given during the progress of the project and ensure the project will complete as targeted.

And also, I would like to thanks all the helps and guidance given by the entire person that participated in completing the project generally.

Lastly, I would like to express my deepest gratitude to my course mates and friends who has also made substantial support in giving idea and key points in producing the project.

Acknowledgement

Here, I would like to express my sincere appreciation to my supervisor, Assoc. Prof. Dr. Selvanathan Narainasamy. He has given the advice and full guidance during my research of this project.

Secondly, I am also grateful and thankful to my moderator Mr. Woo Chaw Seng for his useful advice, guidance and comment for ensuring the project will meet the requirements needed.

I am also very grateful to my group mates Nik Anuar Fathi and Mohd. Faizal, withal the co-operation given during the progress of the project and ensure the project will complete as targeted.

And also, I would like to thanks all the helps and guidance given by the entire person that participated in completing the project generally.

Lastly, I would like to express my deepest gratitude to my course mates and friends who has also made substantial support in giving idea and key points in producing the project.

Table of Content

Abstract

Acknowledgement

Chapter 1 – Introduction

1.1 Project Overview..... 1

1.2 Project Definition..... 2

1.2.1 The Data Mining Paradigm..... 3

1.2.2 Project Limitations and Goal..... 4

1.3 Project Objectives..... 4

1.4 Project Scopes..... 5

1.5 Expected Outcome..... 7

1.6 Project Schedule..... 8

1.7 Report Layout..... 9

Chapter 2 – Literature Review

What is Data Mining?..... 12

2.1 Data Mining Process..... 13

2.1.1 Problem Identification..... 16

2.1.2 Transforming Data into Actionable Results..... 17

2.1.3 Acting to the results..... 20

2.1.4 Measuring the model's effectiveness..... 21

2.2 Data Mining Technique..... 22

2.2.1 Decision Tree..... 23

2.2.1.1 When to use Decision Tree..... 25

2.2.2 Clustering..... 25

2.2.2.1 When to use Clustering..... 27

2.2.3 Neural Networks..... 28

2.2.3.1 When to use Neural Network..... 30

2.3 Others Data Mining Models and Algorithm..... 31

2.3.1 Multivariate Adaptive regression Splines (MARS)..... 31

2.3.2 Rule Induction..... 32

2.3.3 K-nearest neighbor and Memory Based Reasoning (MBR)..... 33

2.3.4 Logistic Regression..... 35

2.3.5 Discriminant Analysis..... 37

2.3.6 Generalized Additive Models (GAM).... 38

2.3.7 Boosting..... 39

2.3.8 Genetic algorithms..... 40

2.4 Database Management Review..... 41

2.4.1 Microsoft SQL Server 2000..... 41

2.4.2 Oracle8i Server..... 43

2.4.3 Microsoft Access 2000..... 45

2.5 Application Programming Language..... 46

2.5.1 Microsoft Visual Basic 6.0..... 46

2.5.2 Java..... 48

2.6 Data Mining Tools..... 49

2.6.1 Weka 3 - Machine Learning Software in Java.... 49

2.6.2 Clementine Data Mining Solution..... 51

Chapter 3 – Methodology

3.1 Methodology..... 53

Chapter 4 – System Analysis and Requirements

4.1 System Analysis..... 59

4.2 Requirements Analysis and Specification..... 61

4.2.1 Functional Requirements..... 61

4.2.2 Non-functional Requirements..... 63

4.3 System Development Tools..... 65

4.3.1 Clementine..... 65

4.3.2 Visual basic 6.0..... 67

4.4 System Requirements..... 68

Chapter 5 – System Design

5.1 Data Flow Diagram..... 70

5.1.1 Context Diagram..... 71

5.2 System Functionality Design..... 72

5.2.1 Modeling Functionality..... 73

5.2.2 User Interface Functionality..... 74

5.3 Database Design..... 74

Chapter 6 – System Implementation

6.1 Platform Development..... 75

6.1.1 Operating System Implementation..... 75

6.1.2 Data Preparation and Implementation..... 77

6.1.3 Clementine Data Mining Solution and Exceed... 78

6.1.4 Visual Studio – Visual C++ 6.0..... 78

6.1.5 Modeling and Exporting Clementine Models..... 79

6.2 Standard and Procedure to Write a Code..... 80

6.2.1 Coding..... 80

Chapter 7 – System Testing

7.1 Unit Testing..... 82

7.1.1 Source Code Examining..... 83

7.1.2 Test Cases..... 84

7.1.3 Data Reliability Testing..... 84

7.1.4 User Testing..... 85

7.2 Integration Testing..... 86

Chapter 8 – System Evaluation

8.1 System Strength..... 87

8.2 System Limitation..... 88

8.3 Future Enhancement..... 89

8.4 Problem Encountered..... 90

8.5 Objectives Achieved..... 91

Chapter 9 – Conclusion..... 93

References..... 95

List of Table

Table 2.3: Time line chart shows score set and model set of data mining model 22

Table 2.4: Training table for decision tree..... 23

Table 6.1: File model generated from Clementine and its dependencies.....81

List of Figure

Figure 1.1: Gantt chart..... 8

Figure 2.1: Data Mining Process..... 16

Figure 2.2: Transforming Data into Actionable Result..... 18

Figure 2.5: Decision Tree..... 24

Figure 2.6: Example graph on Clustering..... 26

Figure 2.7: Neural Networks Structure..... 29

Figure 2.8: K-nearest neighbor..... 34

Figure 3.1: Waterfall Model..... 55

Figure 3.2: Waterfall Model for database analysis using Data Mining Techniques..... 58

Figure 5.1: Context Diagram..... 71

Figure 5.2: System functionality for generating the model.....72

Figure 5.3: System functionality for generating the graphical user interface.....73

Chapter 1 - Introduction

1.1 Project Overview

The era of information technology sets the trend of business to rely on the importance of data and information. Data has become the most valuable resource for a company besides implementing business solutions and automation, which is crucial in maximizing productivity and profit. Techniques that extract the essence of business data gives the advantages in gaining more profit to those who practice good data management rather than others who do not take the opportunity provided. Such cases usually will lead to more problems such as data overwhelming and information exhaustion.

One solution offered by the technology is data mining. Data mining gives its practitioner the step ahead of their competitors where the business is trying to reach their customers and fulfill their needs. This develops a relation called the customer-business relationship. Some of the benefits offered by data mining are target marketing, customer retention prediction and customer profiling, which will be discussed further more.

Hopefully this thesis can help businesses to practice data driven marketing where information is extracted from the data warehouse and its objective is to maximize business revenue while minimizing cost. Hopefully, the customer will experience some sort of personalized catering of their needs or business recognition and customers will be less annoyed by ridiculous offers made by.

1.2 Project Definition

Medical Diagnosis using data mining techniques is a tool using the data mining techniques in prescribing the appropriate drug prescription for the patient that are suffering from the same illnesses.

The system will assist in giving the appropriate drug prescriptions since different patient will interact with different drugs based on their conditions. This means that different persons or patient suffer the same illnesses and symptoms but will interact or respond to different doses of drug given based on certain conditions. The conditions of the patient here are the sex, age, blood pressure, the cholesterol level and other symptoms related to the illness.

Medical Diagnosis using data mining techniques will also figured out which drug will be appropriate for the future patient based on the patterns or the results of the data mining techniques done to the previous data on the patient. This will helps the medical experts to predict the more suitable drug prescriptions to different patients.

1.2.1 The Data Mining Paradigm

Data mining is considered as the identification of information nuggets or decision-making knowledge form large databases. Data mining solutions are becoming more and more popular in our market, which means that the data mining solutions do play an important role in complementing today's business decision support system. Despite this, statistical methods are still widely used to predict customer behavior or other decisions. Statistical methods have been used for targeting marketing efforts for some time now and has proven its effectiveness. This thesis intends differentiate statistical method s and data mining techniques as well as to give alternative solution for businesses, which is easier to use, faster and cost effective.

1.2.2 Project Limitations and Goal

This thesis intends to provide medical information especially in prescribing the drug prescriptions and the patient behavioral. It helps medical experts to understand their patient responds based on their certain condition. This behavioral understanding is the essence of targeted marketing or database driven marketing. Only suggests another method in giving the more appropriate drug prescriptions, not to replace the existing data mining applications. But if this project is a success, more research can be made to provide new reliable and interesting system.

The system had a limitation in prescribing the drugs for a patient that suffering from certain illnesses, because if there are various illnesses, it will take too many parameters to be input in the algorithm.

1.3 Project Objectives

In general, the objectives of this project is to build an intelligent system that can help and assists the medical experts or data analyzers in dealing with prescribing drug for the patient so that can serve even better.

The specific objectives of this thesis are as follows: -

1. Implement a system, which facilitate the process of medical diagnosis in this case the drug prescriptions by using technique derived from data mining activities.
2. The system is able in giving the right doses of drug based on the condition of the patient and also able to prescribe the drug for the future patient.
3. System would be able to transform all the medical data into more valuable information in presenting the patterns of data using the data mining techniques.
4. To achieve and represent the medical information into more meaningful knowledge in giving predictions and assists medical experts in term of prescribing the appropriate drug for their patient.
5. To apply the concept of artificial neural network (ANN) in data mining.

1.4 Project Scopes

As stated before, this system is to provide a method of customer profiling. This system is intended for the marketing team to analyze the data. The application

of this system so assumptions made to limit the scope and capabilities of this system.

1. Assuming that all the data in the databases had been converted to ASCII's files or "Flat files".
2. The output of data mining techniques such as the Predictive Models that gives the patterns of data, and the user has to interact with the information. Data mining just given the valuable information fro the user to interact with it.
3. Some sort of survey/ case study done where questionnaire or any particular data analysis have been made. Realize that data is correct where noise reception must be considered.
4. The use of other methods to support that overall decision support system is highly recommended rather than using this system as a stand-alone product.
5. Will use some sort of user control over the system. Not a standalone system.

1.5 Expected Outcome

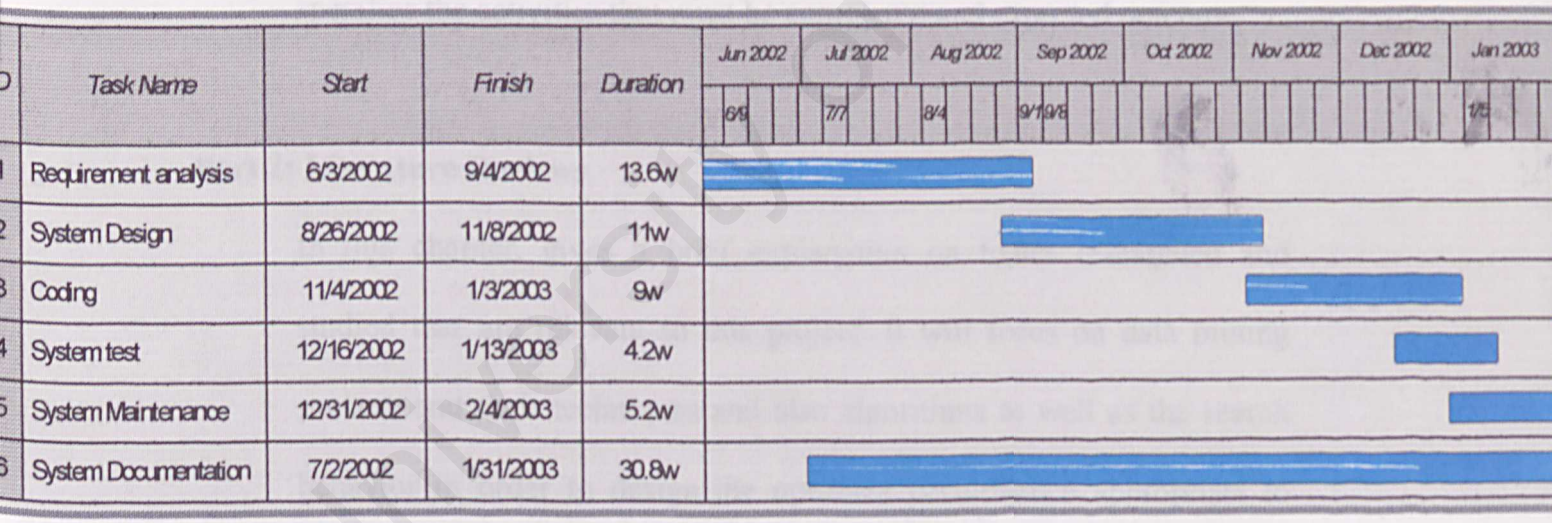
Before coming out with the expected outcome of this project, there a few factor that need to be considered such as amount of time to compare the project and the resources available. Below is some of the expected outcome of the project.

- Implementing the system that can assist the medical experts in prescribing the drug for the patient.
- Provide the pattern of the medical information that has been analyzing from the medical data and then makes prediction based on the pattern or knowledge.
- Able to measure the patterns of data and then convert it into more meaningful information and knowledge.
- The final implementation should allow the future enhancement as well as additional modules to increase the functionality of the system.

1.6 Project Schedule

Project scheduling plays an important role in planning and developing the entire project. It specifies all the activities involve in the project development and the duration of time for each activity to successfully implement in the project.

Figure shows an example of Gantt chart on my project schedule. This is the project planning for the whole system to take place.



	Task Name	Start	Finish	Duration
1	Requirement analysis	6/3/2002	9/4/2002	13.6w
2	System Design	8/26/2002	11/8/2002	11w
3	Coding	11/4/2002	1/3/2003	9w
4	System test	12/16/2002	1/13/2003	4.2w
5	System Maintenance	12/31/2002	2/4/2003	5.2w
6	System Documentation	7/2/2002	1/31/2003	30.8w

Figure 1.1: Gantt chart

1.7 Report Layout

The purpose of this thesis layout or the report layout is to give an overall overview of the major contents, which will be included and involved during the development of this project.

Part 1: Introduction

To gives an overview of the project, which included the project definition, project objectives, project scopes, project schedule that specifies the activities that must be completed and expected outcome.

Part 2: Literature Reviews

In this chapter, gives a brief explanation on topics researched and studied that are relevant to this project. It will focus on data mining process and their techniques and also algorithms as well as the search behavior in order to design the optimize performance appropriate to them.

Part 3: Methodology

Emphasized on methodology and chosen model to develop the project as well as the development tools. This chapter focuses on the model

chosen to develop and explain the benefits and why it is chosen as miles stones. It also shows on how data mining techniques interact with the model.

Part 4: System Analysis and Requirements

Emphasized and explains how the requirements for this project were required and the analysis of the results. Besides that, it also analysts the development tools that are available and then choose the best tools of software to develop the system.

Part 5: System Design

This chapter explains the conceptual and technical design process of the system. It will include the database and reports.

Part 6: System Implementation

This chapter emphasis on the implementation tools use to create the whole system. In this cases, Clementine Data Mining Solution, Visual C++ and Microsoft Foundation Classes (MFC) used to develop the system.

Part 7: System Testing

This chapter explains the testing done to the system which includes the source code examining, test cases, unit testing and testing the reliability of the output resulted from the system.

Part 8: System Evaluation

Explanations about the evaluation of the overall system including the system strength and limitations, the future enhancement can be done. The chapter also emphasis on the other issue like the problem faced when developing the system.

Part 9: Conclusion

The chapter is an overview of the system and conclusion about the project for the under-graduated student. This section allows the student to suggest any suggestion or approach in order to enhance the efficiency of the project for the future student.

Chapter 2 – Literature Review

The main purpose of literature review is to find any related information about the system that we intend to develop. By gathering all the information, a better product can be developed. Selecting the better software, tools, and approaches is important for a best outcome and also meet the specification requirements. Without this analysis, we would not be able to identify the strengths and weaknesses of each tool.

Data mining definition that has been stated as the main techniques will be used in most of the discussion in material that have been reviewed. It will be a base line throughout this project, but will not show how the define it and why it is defined like that. In the following discussion, we will concert more on the process of data mining.

What is Data Mining?

By referring the book of Mastering Data Mining: The Art and Science of Customer Relationship Management by Michael J.A. berry and Gordon S. Linoff, the definition of data mining stated as

> *Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules.*

In other words, data mining is the semi-automatic discovery of patterns, associations, changes, anomalies, rules, and statistically significant structures and events in data. That is, data mining attempts to extract knowledge from data.

Data mining is about finding interest structures in data, which may be interpreted as knowledge about the data or may be used to predict events related to the data. These structures take the form of patterns, which are concise descriptions of the data set. Data mining makes the exploration and exploitation of large databases easy, convenient, and practical for those who have data but not years of training in statistic or data analysis.

The "knowledge" extracted by a data mining algorithm can be many forms and any use. It can be in the form at a set of rules, a decision tree, a regression models, or a set of associations, among many other possibilities. It may be used to produce summaries of data or to get insight into previously unknown correlations. It also may be used to predict events related to the data – for example, missing values, records for which some information is not known, and so forth.

Data mining differs from traditional statistic in several ways: formal statistic inference is assumption driven in the sense that a hypothesis is formed and validated against the data. Data mining in contrast is discovery driven in the sense that patterns and hypothesis are automatically extracted from data. Serial other way, data mining is data driven, while statistic is human driven. The branch of statistic that data mining resembles most is exploratory data analysis, although this field, like most of the rest at statistic, has been focused on data sets for smaller than most that use the target of data mining researchers.

Data mining also differs from traditional statistics in that sometimes the goal is to extract qualitative models which can easily be translated into logical rules or

visual representations; in this sense data mining is human centered and is sometimes coupled with human-computer interfaces research.

Data mining is a step in the data mining process, which is an interactive, semi-automated process which with new data. Results of the data mining process may be in gusts, rules, or predictive models.

2.1 Data Mining Process

Through some references, concluded that data mining process behaved like software lifecycle, cycling through four main processes. Here are the four processes: -

- Business problem identification
- Data transformation
- Act to the result
- Measurement

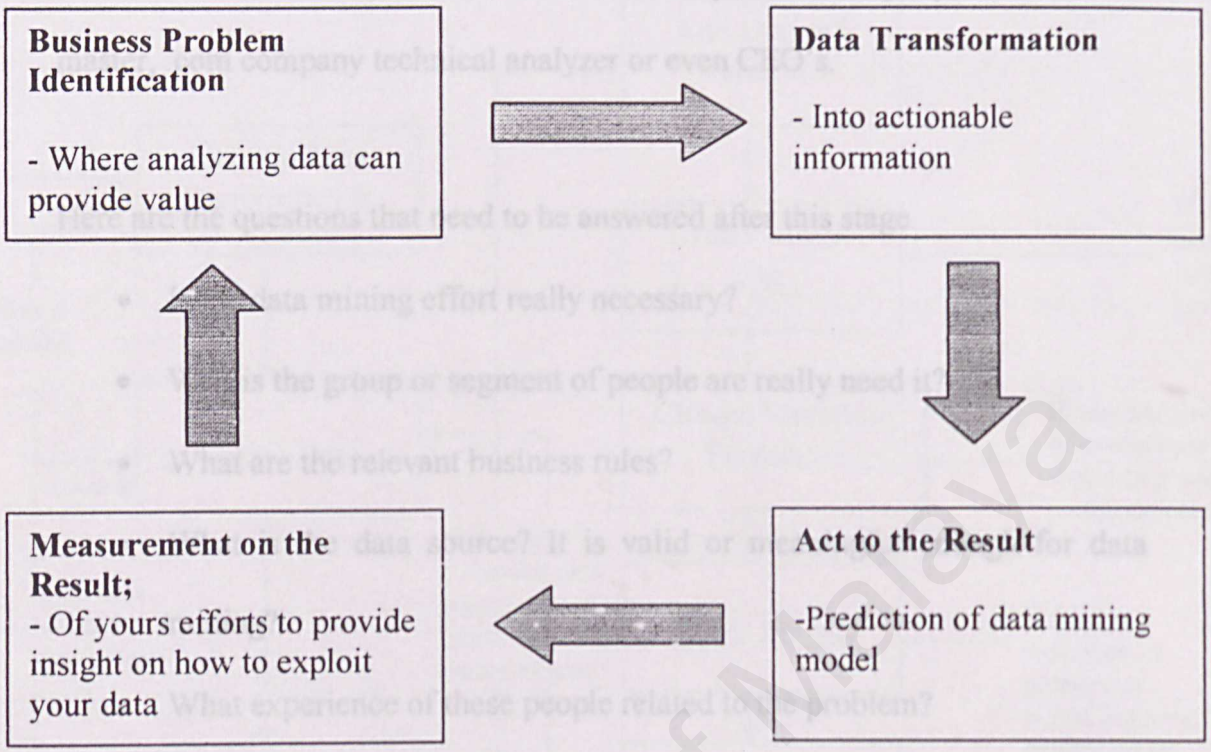


Figure 2.1: Data Mining Process

2.1.1 Problem Identification

In business, people who understanding the business are always considered as domain experts. Meaning, these people who we need to ask for the real problem need to be solved. In this stage, communication skill needed and the communication may become a great challenge.

In our case, technically, who is the domain expert? These people can be web master, .com company technical analyzer or even CEO's.

Here are the questions that need to be answered after this stage.

- Is the data mining effort really necessary?
- Who is the group or segment of people are really need it?
- What are the relevant business rules?
- What is the data source? It is valid or meaningful enough for data mining?
- What experience of these people related to the problem?

2.1.2 Transforming Data into Actionable Results

The most important of data mining is to transfer data into actionable results. The goal is to build a data mining models. Illustrated are the basic step taken to transforms data into actionable results.

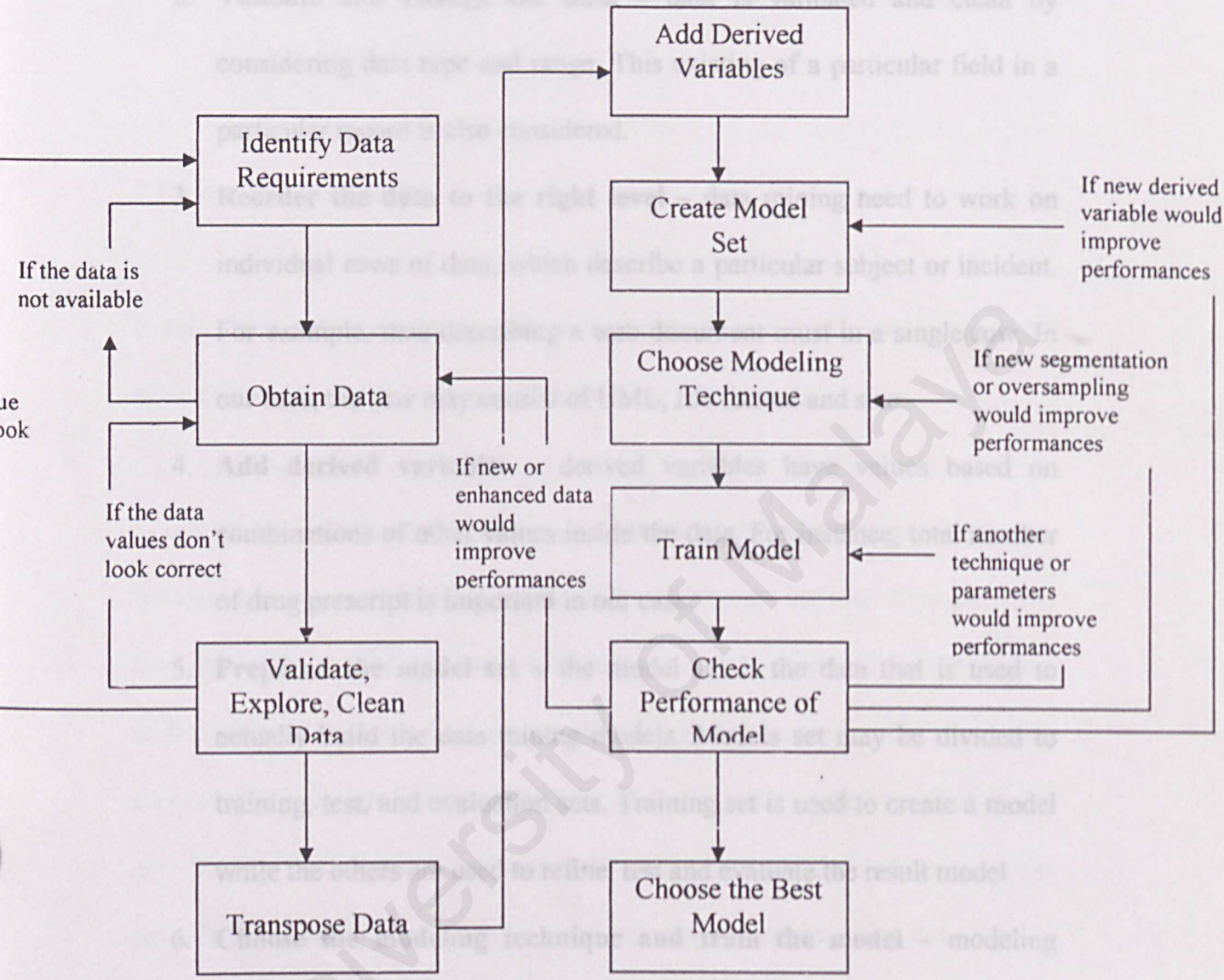


Figure 2.2: Transforming Data into Actionable Result

1. **Identify and obtain data** - the first steps in the modeling process are identifying and obtain the right data. This data obtained should meet the requirement for solving our problem.

2. **Validate and change the data** – data is validated and clean by considering data type and range. This existing of a particular field in a particular record is also considered.
3. **Reorder the data to the right level** – data mining need to work on individual rows of data, which describe a particular subject or incident. For example, data describing a web document must in a single row. In our case, the row may consist of UML, file format and size.
4. **Add derived variables** – derived variables have values based on combinations of other values inside the data. For instance, total number of drug prescript is important in our case.
5. **Prepared the model set** – the model set is the data that is used to actually build the data mining models. Models set may be divided to training, test, and evaluation sets. Training set is used to create a model while the others are used to refine, test and evaluate the result model.
6. **Choose the modeling technique and train the model** – modeling technique like decision tree, clustering, and neural network will be used to train the model by using the model set data as input. These three techniques will be discussed later in this chapter.
7. **Check performance of the model** – different ways of evaluating the outcome and cumulative gain chart is two ways to compare the

performance of different models. In order to do so, evaluate set (an unused set of data) is used.

2.1.3 Acting to the results

In this stage, result can be seen from the prediction of data mining model. All we used to do is to solve our problem by using three results.

In our context, problem can be gaining some business or technical knowledge. For example, we can identify the user access pattern on website. Also the result may just used to make a decision whatever we need to drop a file or not depend on the accesses of users. This is considered as one-time results. The result may also use as prediction following the time line – periodic predictions. The model may be used to determine the number of access to particular web page or site periodically. Lastly, the result may be used as feedback on a data mining process where the data need to be fixed, validated, and clean.

2.1.4 Measuring the model's effectiveness

In previous section, evaluation by using confusion matrix and cumulative gains chart has been introduced. It is a way to evaluate but not measure. Our measure here is to compare the real and actual world with the predicted results. Let say, as an example here, a number of patients has been diagnosed and given the appropriate drug for their symptoms in certain time duration. After the prediction, we will measure accuracy by compare the figure to the real one.

In future discussion to differential the evaluation stage and measurement stage, Figure is collected from references. This figure show six months of historical data is used as a training set- data that used to create model. And **P** is used as evaluation set- data set that used to evaluate model. In the last time line, the **P** here is not an evaluation set, it may be considered as measurement set or predicted set. This set is the results of the prediction done by data mining model. We call it measurement set because me may have an actual set to measure the accuracy of the model.

JAN	FEB	MAR	APR	MAY	JUNE	JULY	AUG	SEPT	OCT	NOV
Model Set										
6	5	4	3	2	1		P			
	6	5	4	3	2	1		P		
Score Set										
		6	5	4	3	2	1		P	

Table 2.3: Time line chart shows score set and model set of data mining model

2.2 Data Mining Technique

Data mining technique determine how the cases for a data-mining model are analyzed. Data mining model technique provide the decision-making capabilities that needed to be clarify, segment, associate, and analyze data for the processing of data mining columns that provide predictive, variance, or probability information about the case set.

Here, we will discuss 3 categories of data mining algorithms, which are

- Decision Tree
- Clustering
- Neural Networks

2.2.1 Decision Tree

Decision tree is a very well known algorithm use in one form or another by almost all commercially available data mining tools. Decision tree algorithms are recommended for predictive task that require a classification-oriented. In general, this technique builds a tree that will predict the value of a column based on the remaining columns in the training set. Therefore, each node in the tree represents a particular case for a column.

Decision tree will make the decision on where to place this node. And for a node at a difficult depth than its siblings may represent different cases of each column. For instance, consider the following training table.

Share Files	User Scanner	Infected Before	Risk
Yes	Yes	No	High
Yes	No	No	High
No	No	Yes	Medium
Yes	Yes	Yes	Low
Yes	Yes	No	High
No	Yes	No	Low
Yes	No	Yes	High

Table 2.4: Training table for decision tree

For this training data, the following decision tree may be produced

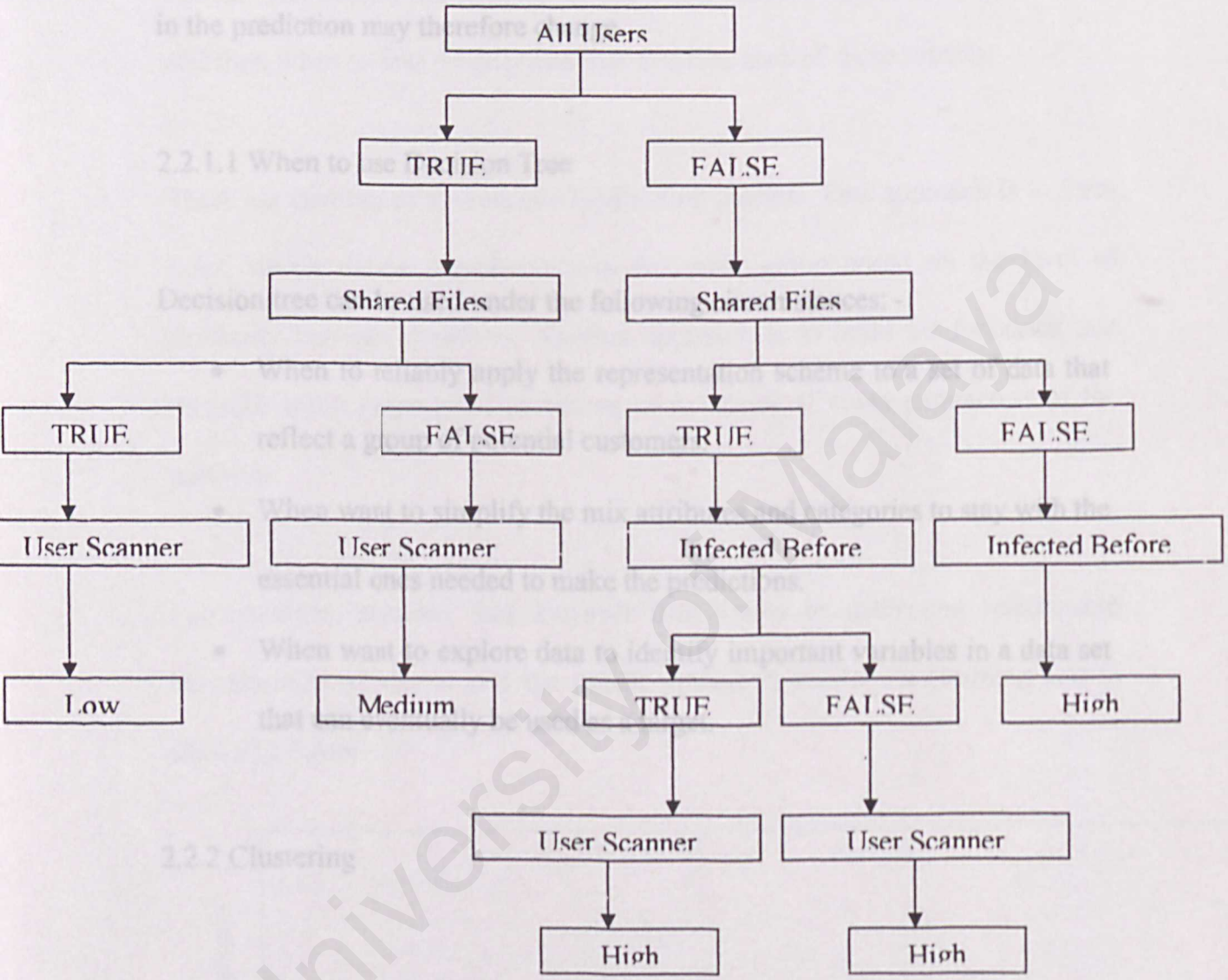


Figure 2.5: Decision Tree

As conclusion on Figure, for users that share file, the most important factor fro determining their risk of computer virus effect is User Scanner. For user who didn't share files, the most important factor is Infected Before. This shown that

a column may be used at more than one location in the tree, and its importance in the prediction may therefore change.

2.2.1.1 When to use Decision Tree

Decision tree can be used under the following circumstances: -

- When to reliably apply the representation scheme to a set of data that reflect a group of potential customers.
- When want to simplify the mix attributes and categories to stay with the essential ones needed to make the predictions.
- When want to explore data to identify important variables in a data set that can eventually be used as a target.

2.2.2 Clustering

Clustering in data mining context is defined as: -

"Clustering algorithm is an expectation method. It uses iterative refinement techniques to group records into clusters that similar, predictable characteristic"

Clustering according to similarity is a very powerful technique, the key to it being able to translate some intuitive measure of similarity in to a quantitative

measure. When leaving is unsupervised then the system has to discover its own classes. The system has to discover subsets of related objects in the training set and then it has to find descriptions that describe each of these subsets.

There are number of approaches for forming clusters. One approach is to form rules, which dictate membership in the same group based on the level of similarity between members. Another approach is to build set functions that measure some property of partitions as functions of some parameters of the partition.

For example, suppose that a travel firm wants to determine relationship between age of visitor and the places visited. A model set (training set) is shown in figure.

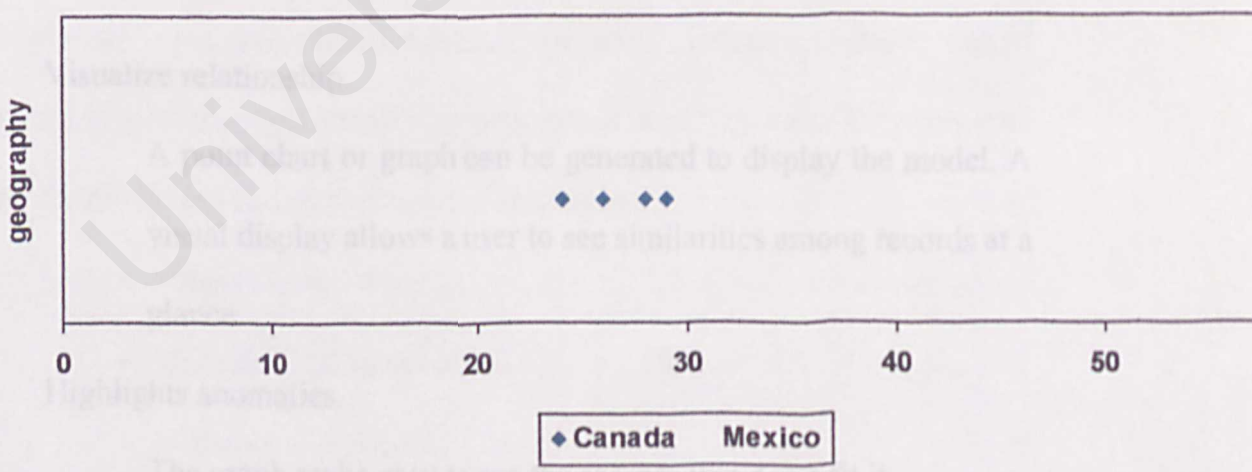


Figure 2.6: Example graph on Clustering

By this example, the cluster of data can be observed. This training set is only consists of two fields- two dimension. For data with higher dimensions, plotting the data in this manner may not be convenient or may be impossible. The clustering techniques in data mining tools will automatically find the way to cluster for data with higher dimensions.

2.2.2.1 When to use Clustering

Clustering is the best choice of algorithm when there is a large quantity of data that has a high degree of logical structure and many variables.

Clustering can be used to: -

i. Visualize relationship.

   A point chart or graph can be generated to display the model. A visual display allows a user to see similarities among records at a glance.

ii. Highlights anomalies.

   The graph make easy to see the records that don't fit it.

iii. Create samples for other data-mining efforts.

To targets this group, instead of choosing known categories, clustering algorithm is applied and then cases present in more meaningful nodes are chosen.

2.2.3 Neural Networks

Neural networks are among the most complicated of the classification and regression algorithms. Although training a neural network can be time consuming, a trained neural network can speedily make predictions for new cases.

Neural networks have brood applicability to real world business problems and have already seen successfully applied in many industries. Since neural networks are the best at identifying patterns or trend in data, they are well suited for prediction or forecasting needs including:

- Stakes forecasting
- Industrial process control
- Customer research
- Data validation

- Risk management
- Target marketing, etc.

The structure of a neural network looks something like the following figure.

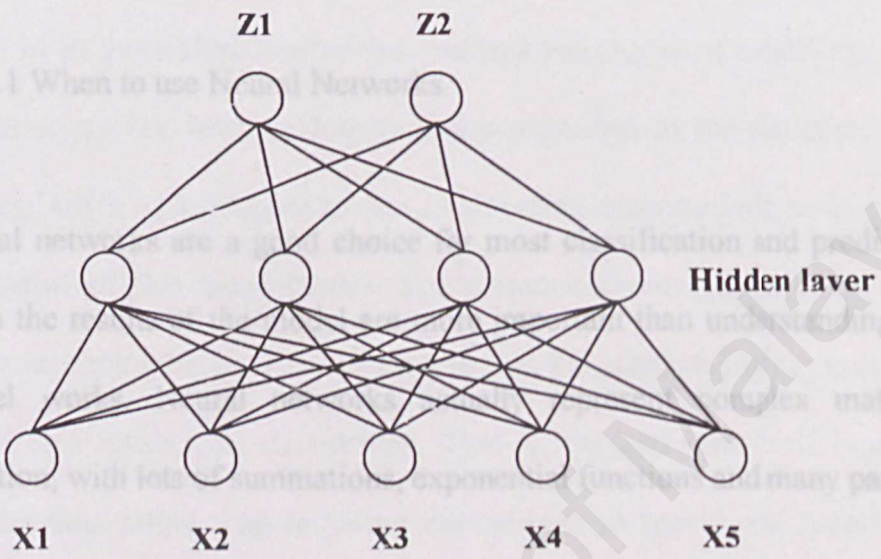


Figure 2.7: Neural Networks Structure

The bottom layer represents the input layer (source layer), in this case with 5 inputs labeled as X1 through X5. In the middle is something called the hidden layer, with a variables number of nodes. It in the hidden layer that performs much of the work of the networks. The output layer in this case has two nodes, Z1 and Z2 representing output values we are trying to determine from the inputs. For example, predict sales (output) based on the past sale, price and season (input). IN the medical cases, such as the drug prescriptions, the input

will much likely be the symptoms, age, sex, blood pressure, cholesterol level for certain illnesses. The outputs are certainly the appropriate drug prescriptions for the diagnosed patient. And the patterns (output) will help determine the appropriate drug for the other patient that suffers the same symptoms.

2.2.3.1 When to use Neural Networks.

Neural networks are a good choice for most classification and prediction task when the results of the model are more important than understanding how the model works. Neural networks actually represent complex mathematical equation, with lots of summations, exponential functions and many parameters.

Neural networks does not work well when there are many hundreds or thousand of input features. Large numbers of features make it more difficult to find patterns and can results training phases that never converge in good solutions.

2.3 Others Data Mining Models and Algorithm

There are also others method and techniques in implementing data mining used to mine data. Most of the models and algorithm discussed in this section can be taught of as generalizations of the standard workhorse of modeling, the linear regression model. Much effort has been expended in the statistics, computer science, artificial intelligence, and engineering communities to overcome the limitations of this basic model. The common characteristic of many of the newer technologies we will consider is that the patterns-finding mechanism is data-driven rather that user-driven. That is, the software itself based on the existing data rather than requiring the modeler to specify the functional form and interactions finds the relationships inductively. Consequently, we will need a variety of tools and technologies in order to find the best possible model.

2.3.1 Multivariate Adaptive regression Splines (MARS)

The basic idea of MARS is quite simple, while the algorithm itself is rather involved. Very briefly, the Classification And Regression Trees (CART) disadvantages are taken care of by: -

- Replacing the continuous branching at a node with a continuous transition modeled by a pair of straight lines. At the end of the model-building process, the straight lines at each node are replaced with a very smooth function called a spline.
- Not requiring that new splits be dependent on previous splits.

Unfortunately, this means MARS loses the tree structure of CART and cannot produce rules. On the other hand, a mar automatically finds and lists the most important predictor variables as well as the interactions among predictor variables. MARS also plots the dependence of the response on each predictor. The result is an automatic non-linear step-wise regression tool.

MARS, like most neural net and decision tree algorithms, has a tendency to over fit the training data. This can provide good prediction on the test set. Second, there are various tuning parameters in the algorithm itself that can cross validation.

2.3.2 Rule Induction

Rule induction is a method for deriving a set of rules to classify cases. Although decision tree can produce a set of rules, rule induction methods generate a set of

independent rules, which do not necessarily (and are unlikely to) form a tree. Because the rule inducers is not forcing splits at each level, and can look ahead, it may be able to find different and sometimes better patterns for classification. Unlike trees, the rules generated may not cover all possible situations. Also unlike trees, rules may sometimes conflict in their predictions, in which case it is necessary to choose which rule to follow. One common method to resolve conflicts is to assign a confidence to rules and use the one in which you are most confident. Alternatively, if more than two rules conflict, you may let them vote; perhaps weighting their votes by the confidence you have in each rule.

2.3.3 K-nearest neighbor and Memory Based Reasoning (MBR)

When trying to solve new problems, people often look at solutions to similar problems that they have previously solved. K-nearest neighbor (k-NN) is a classification technique that uses a version of this same method. It decides in which class to place a new case by examining some number- the "k" in k-nearest neighbor –of the similar cases or neighbors. It counts the numbers of cases for each class, and assigns the new case to the same class to which most of its neighbor belongs.

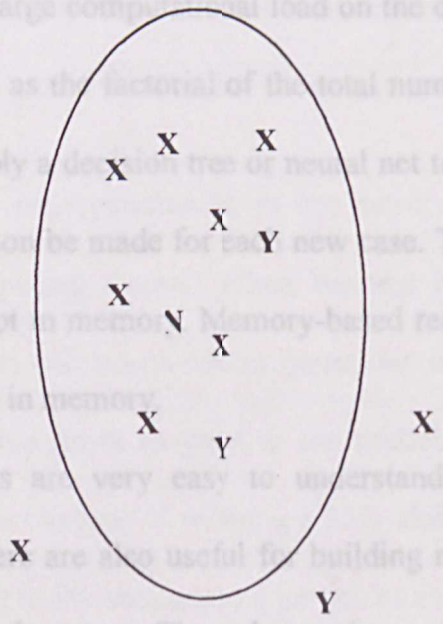


Figure 2.8: K-nearest neighbor

N is a new case. It would be assigned to the class X because the seven X's within the eclipse outnumber the two Y's.

The first thing must do to apply k-NN is to find a measure of distance between attributes in the data and then calculate it. While this is easy for numeric data, categorical variables need special handling. For example, what is the distance between blue and green? You must then have a way of summing the distance measures for the attributes. Once you can calculate the distance between cases, decide how large a neighborhood in which to do the comparisons, and also decide how to count the neighbors themselves.

K-NN puts a large computational load on the computer because the calculation time increases as the factorial of the total number of points. While it's a rapid process to apply a decision tree or neural net to a new case, k-NN requires that anew calculation be made for each new case. To speed up k-NN, frequently all the data is kept in memory. Memory-based reasoning usually refers to a k-NN classifier kept in memory.

K-NN models are very easy to understand when they're a few predictor variables. There are also useful for building models that involve non-standard data types, such as text. The only requirement for being able to include a data type is the existence of an appropriate metric.

2.3.4 Logistic Regression

Logistic regression is a generalization of linear regression. It is used primarily for predicting binary variables and occasionally multi-class variables. Because the response variable is discrete, it cannot be modeled directly by linear regression. Therefore, rather than predict whether the event itself will occur, we build the model to predict the logarithm of the odds of its occurrence. This logarithm is called the log odds or the logit transformation.

The odds ratio:

$$\frac{\text{Probability of an event occurring}}{\text{Probability of an event not occurring}}$$

Has the same interpretation as in the more casual use of odds in games of chances or sporting events. When we say that the odds are 3 to 1 that a particular team will win a soccer game, we mean that the probability of their winning is three times as great as the probability their losing. So we believe they have 75% chances of winning a 25% chance of losing. Similar technology can be applied to the chances to a particular customer replying to a mailing.

While logistic regression is a powerful modeling tool, it assumes that the response variable is linear in the coefficient of the predictor variable. Furthermore, the modeler, based on experience on data and data analysis, must choose the right inputs and specify their functional relationship to the response variable. Additionally the modeler must explicitly add terms for any interactions. It is up to the model builder to search for the right variables, find their correct expression, and account for their possible interactions. Doing this effectively requires a great skill and experience on the part of the analyst.

2.3.5 Discriminant Analysis

Discriminant analysis is the oldest mathematical classification technique; it finds hyper-planes that separate the classes. The resultant model is very easy to interpret because all the user has to do is determine on which side of the line a point falls. Training is simple and scalable. The technique is very sensitive to patterns in the data. It is used very often in certain disciplines such as medicine, the social sciences, and field biology.

Discriminant analysis is not popular in data mining, however, for three main reasons. First, it assumes that all of the predictor variables are normally distributed, which may not be the case. Second, unordered categorical predictor variables cannot be used at all. Third, the boundaries separate the class all linear forms, but sometimes the data just can't be separated that way.

Recent versions of discriminant analysis addresses some of these problem by allowing the boundaries to be quadratic as well as linear, which significantly increases the sensitivity in certain cases. There are also techniques that allow the normality assumption to be placed with an estimate of the real distribution.

Ordered categorical data can be modeled by forming the histogram from the bins defined by the categorical variables.

2.3.6 Generalized Additive Models (GAM)

There is a class of models extending both linear and logistic regression, known as generalized additive models or GAM. They are called additive because we assume that the model can be written as the sum of possibly non-linear functions, one for each predictor. GAM can be used either for regression or for classification of a binary response. The response variable can be virtually any function of the predictors as long as there are not discontinuous steps. For example, suppose that payment delinquency is a rather complicated function of income where the probability of delinquency initially declines as incomes increases. It then turns around and starts to increase again for moderate income, finally peaking before coming down again for a higher income card-holders. In such a case, a linear model may fail to see any relationship between income and delinquency due to the non-linear behavior. GAM, using computer power in place of theory or knowledge of the functional form, will produce smooth curve, summarizing the relationship as describe above. The most common estimation produce is backfitting. Instead of estimating large numbers of

parameters as neural nets do, GAM goes step further and estimates a value of the output for each value of the input –one point, one estimate. As with the neural net, GAM generates a curve automatically, choosing the amount of complexity based on the data.

2.3.7 Boosting

If building a model using a sample of data, and then build a new model using the same algorithm but on the different sample, might get a different result. After validating the two models, we could choose the one that best met the objectives. Even the better result might be achieved if built several models and let it vote, making a prediction based on what the majority recommended. Of course, any interpretability of the prediction would be lost, but the improved results might be worth it.

Basically, boosting takes multiple random samples from the data and builds a classification model for each. The training set is changed based on the result of the previous models. The final classification is the class assigned most often by the models. The exact algorithms for boosting have evolved from the original, but the underlying idea is the same.

2.3.8 Genetic algorithms

Genetic algorithms are not used to find patterns per se, but rather to guide the learning process of data mining algorithms such as a method for performing a guided search for good models in the solutions space.

They are called genetic algorithms because they loosely follow the pattern of biological evolution in which the members of one generation compete to pass on their characteristics to the next generation, until the best is found. The information to be passed on is contained in "chromosomes", which contain the parameters for building the model.

For example, in building a neural net, genetic algorithms can replace backpropagation as a way to adjust the weights. The chromosomes in this case would contain the weights. Alternatively, genetic algorithms might be used to find the best architecture, and the chromosomes would contain the number of hidden layers and the number of nodes in each layer.

2.4 Database Management Review

There are a few database management will be reviewed here, such as SQL 2000, Oracle8i, and Microsoft Access. Each has their weaknesses and strong points.

2.4.1 Microsoft SQL Server 2000

Microsoft SQL Server 2000 provides major enhancement to the current SQL Server product. Numerous changes have been made across the product, including the Relational Engine, the Storage Engine, Administration and tools, replication, the Analysis Service (OLAP Services and Data Mining), English Query, Full-text Search, Integration with Windows® 2000 (with Active Directory™, Address Windowing Extensions, and Windows 2000 Data Center), and Meta Data Services (including the Repository Engine).

SQL Server 2000 includes many new features that extend its capabilities as a high performance relational database system with a rich development environment. SQL Server 2000 is a perfect example of an n-tier system. The user can manipulate the data directly from the client-side. It also maintains

referential integrity and security and ensures that operation can be recovered in the event of numerous types of failure. SQL Server can control the access for the type of information that can be retrieved by the user.

SQL Server 2000 makes giant strides in performance, reliability, and scalability, giving organizations many opportunities to create intelligent real-world business solutions. By giving a need for more simplified and cost-saving features, organizations inspired the following innovations in SQL Server 2000:-

- Scalable from laptop to multiprocessor cluster.
- Dynamic row-level locking.
- Dynamic self-management.
- Wide array of replications options.
- SQL Server desktop.
- Integrated OLAP Services.
- Data transmission services.
- Microsoft English Query.
- Microsoft Repository.
- Integrated with Microsoft Office 2000 and Microsoft Visual Studio.

2.4.2 Oracle8i Server

The entire Oracle8i database is designed for power and ease in Internet development. All Oracle8i database are compatible and available on many different platforms, so database applications can scale from handheld to laptop to enterprise without changing our applications.

Oracle8i Enterprise e-business database is a powerful database available for driving enterprise e-business applications, online transaction processing applications (OLTP), Query-intensive data warehouses, and high capacity web sites. Oracle8i Enterprise Edition has an unparallel performance and stability that spans from single CPU server through massive server clusters and maintenance.

With the integration of innovative technologies like native support for XML, Java and SQL, the Oracle8i Enterprise Edition is ready to power the e-business. Oracle Database enterprise Edition delivers unprecedented ease-to-use, power, and price or performance. It includes: -

- Scalable, browser-based system management for end-to-end management of your entire e-business.

- Optional system management packs for comprehensive system diagnostic and timing operations.
- Powerful Oracle Partitioning option for effectively managing terabytes of data with a minimum of administrative overhead.
- Strong, cost-effective data security designed for today's web and e-business environments.
- Support for highly available and scalability cluster configurations with the Oracle Parallel Server options.
- Fast and easy storage and management of MS Word documents, Excel files, email, multimedia, XML files, and web pages in the Oracle Internet File Systems.
- Oracle8i Enterprise Edition supports the latest Symmetrical Multi-Processor (SMP) systems and can be extended with the following options: Oracle Advanced Security, Oracle Label Security, Oracle Parallel Server, Oracle Partitioning, Oracle Spatial, Oracle Visual Information Retrieval, Oracle Diagnostic Management Pack, Oracle Tuning Management pack and Oracle Change Management pack.

2.4.3 Microsoft Access 2000

Microsoft Access 2000 simplifies the skill set needed to create simple, useful databases; the improved interface offers more consistency with other Office applications, plus new features that increase productivity. Access 2000, through its support of OLE DB, can act as a front-end to high-end database engines such as Microsoft SQL Server, making Access 2000 databases more scalable than ever before.

Access 2000 can act as a front-end client to corporate-level, back-end databases, such as Microsoft SQL Server. Access can now be used in two ways: as a standalone application for creating databases for individual or department use or as an easy-to-use interface client to a more scalable and robust back-end database that was previously only available to professional database administrators (DBA's). This lowers the bar for creating true client/server applications by allowing and users to take advantage of the ease-to-use of Access combined with the scalability and reliability of Microsoft SQL Server. Regardless of the back-end data source selected, end users will still have the same easy-to-use experience of the most popular desktop database client.

Some of the features that Access provide as:

- Enable web collaboration and improve productivity.
- Seamless integration between our data source and interactive web. Pages make building and sharing an Access database easier.
- Together with the Office 2000 web components to usually analyze data in our web browser.
- Access 2000 allows quickly analyze details and see vital relationship.
- Access 2000 includes built-in SQL Server integration that brings the power of high-end database management to the familiar Access environments.

2.5 Application Programming Language

There are a two programming language will be reviewed here, which are Visual Basic 6.0 and Java.

2.5.1 Microsoft Visual Basic 6.0

Visual Basic is an extremely powerful, full-featured application development tools that exploits the key features of Microsoft Windows. It is easy to use

through a graphical interface. Applications can be built in the short time by using it.

Using integrated visual database tools, advanced database application can be developed to access SQL Server database, Access database or any third party database by using the ODBC, DAO, RDO, or ADO and bind the data to forms and reports that greatly reduces development time.

Besides that, it provides support for Graphical User Interface (GUI) design that helps interface designer to enhance screen design. We use controls to create the user interface of an application, including command buttons, check boxes, combo boxes, text boxes, scroll bars, frames, files, and directory selection boxes, timers, and menu bars.

Microsoft Visual Basic 6.0 is also an interpreted language system, so users could test and debug application on the fly form within developments environments.

2.5.2 Java

Java is a perfect programming language for the Internet. It is platform-independent which means that a single application can run on a computer running Windows 95/98/NT, UNIX, MVS or a Mac OS.

Java also allows the creation of applets, which are small applications designed to be embedded in a web document. Applets can animate graphics, play audio clips and interact with the users through graphical user interface.

The Java programming language and environments is designed to solve a number of problems in modern programming practice. It stated as a part of a larger project to develop advanced software for consumer electronics.

These devices are small, reliable, portable, distributed, and real-time embedded systems.

2.6 Data Mining Tools

2.6.1 Weka 3 - Machine Learning Software in Java

Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License.

Implemented schemes for classification include:

- decision tree inducers
- rule learners
- naive Bayes
- decision tables
- locally weighted regression
- support vector machines
- instance-based learners
- logistic regression

- voted perceptrons
- multi-layer perceptron

Implemented schemes for numeric prediction include:

- linear regression
- model tree generators
- locally weighted regression
- instance-based learners
- decision tables
- multi-layer perceptron

Implemented "meta-schemes" include:

- bagging
- stacking
- boosting
- regression via classification
- classification via regression
- cost sensitive classification

2.6.2 Clementine Data Mining Solution

Clementine is a data mining toolkit. It was design to helps to makes the most profitable use of what is, a greatly underutilized resource-data. Clementine's "visual programming" interface is easy to use but supports rich facilities for manipulating data , experimenting with different combinations of data and of techniques, and for testing hypothesis. This facility helps to make it easy to apply knowledge and expertise to discover interesting and valuable properties of your data.

Clementine can also give more active assistance. It includes advanced modeling or "machine learning" technologies, which extract complex interrelationship and decisions-making rules from the data. These help to automate applications such as prediction, estimation and classification, and can be used to provide "expert" decision support.

Clementine includes a number of machine learning and modeling technologies, including rule induction, neural networks, association rule discovery and clustering. It also includes many facilities to let bring expertise to bear on data, such as: -

- data manipulation – constructing new data items derived from existing ones, and breaking the data down into meaningful subsets.
- browsing and visualization – displaying aspects of the data using interactive graphics.
- statistics – confirming suspected relationship between factors in data.
- hypothesis testing – constructing models on how the data behaves, verifying them

Typically, we will use these facilities to identify a promising set of factors in the data; these can then be fed to the modeling techniques, which will attempts to identify underlying rules and relationships.

Chapter 3 – Methodology

3.1 Methodology

In order to develop this thesis on time, an effective development method had to be chosen. Methodology is classically though out of as set of activities fact analysts, designers and user carry out to develop and implement a system. A suitable methodology helps the author to develop the system on time and increase the quantity or usefulness of the system. The Waterfall Model has been chosen as the development methodology for Medical Diagnosis using AI Techniques.

There are a few reason of why Waterfall Model is chosen for my project. The reasons are: -

- It is widely used, easily understood and implemented in a system development process.
- It enforces disciplined approach to develop a system as documents prepared after each stage will have to be checked and approved.

- It supports good process visibility as each activity produces some kind of deliverable. These deliverables may prove to be useful when the system evolves in the future.
- It enables maintenance to be carried out at each stage due to its interactive nature. Changes can be done during any of the stages by returning to the previous stages. The iteration process may be carried out as many times as needed and this produces a fine system of high quality that meets a user's requirements.

These are also a few reasons of why prototyping is required with waterfall model: -

- It can help to control the thrashing and therefore enhance understanding.
- A prototyping is a partially developed system and decide if it suitable or appropriate for the finished product.
- Prototyping is useful for verification and validation where verification ensures that each function works correctly and validation ensures that the system has implemented the entire requirements in the specification.

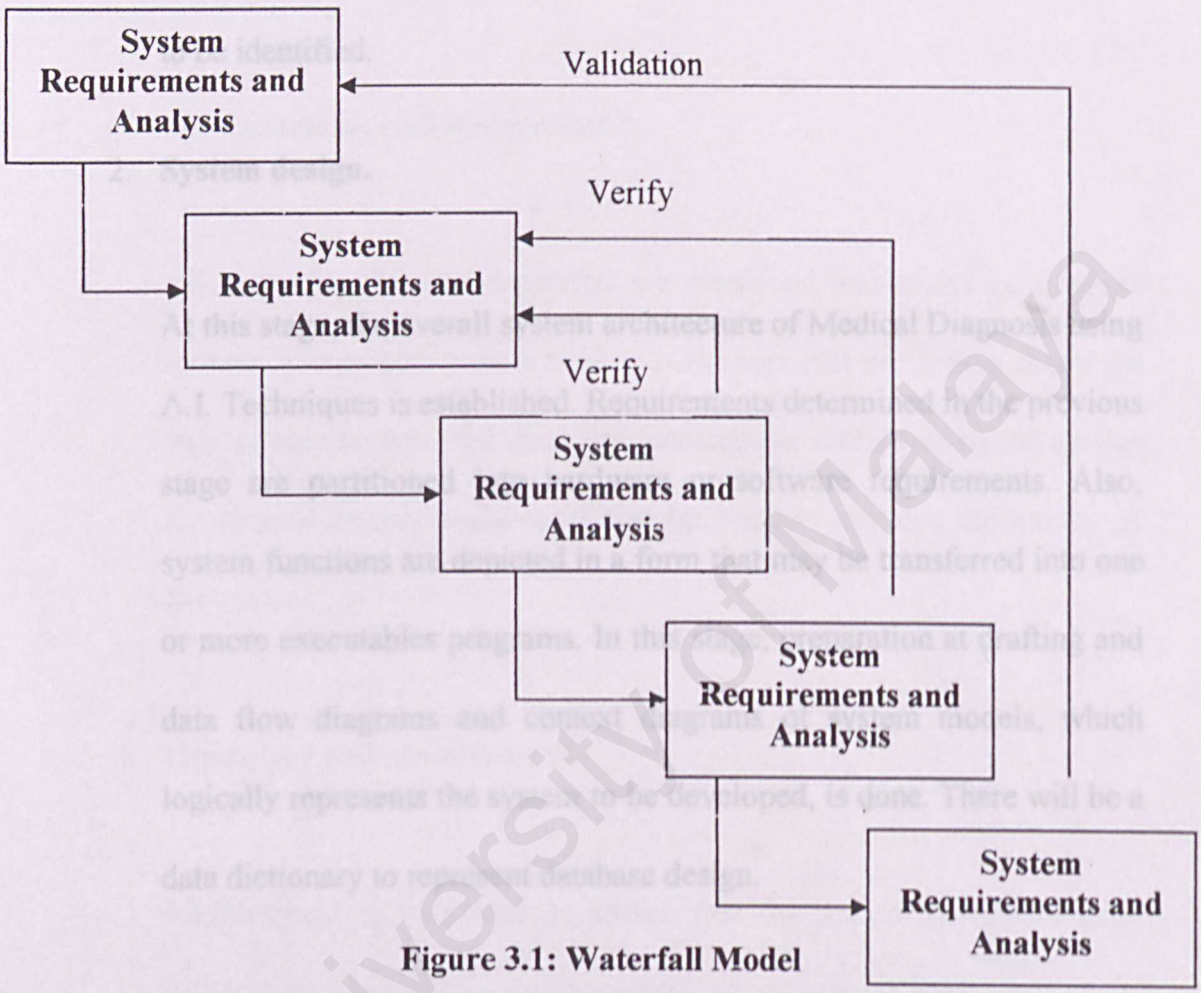


Figure 3.1: Waterfall Model

1. **System analysis and requirements.**

This stage has to be done in chapter 2. All the information regarding this project is gathered. Information is gathered through Internet and reading materials like books, magazines, journals, and newspaper. The main

objectives of this stage are to establish the system's services constraints and goals. In this stage project, requirements, needs and constraints have to be identified.

2. **System design.**

At this stage, the overall system architecture of Medical Diagnosis using A.I. Techniques is established. Requirements determined in the previous stage are partitioned into hardware or software requirements. Also, system functions are depicted in a form that may be transferred into one or more executables programs. In this stage, preparation at drafting and data flow diagrams and context diagrams of system models, which logically represents the system to be developed, is done. There will be a data dictionary to represent database design.

3. **Implementation.**

In this stage, all programs will be coded using the selected programming language or application development tools following the design specify in the System Design. Each function will then be tested to verify that it

is working to its specifications. During this stage, various bugs will be eliminated.

4. **System testing and integration.**

   All the units that are separated are combined and tested as a whole system. The system testing aims to make sure that the system meets the user's requirements and therefore, ensured the usefulness of the system developed. Enhancements will also be made to improve the quality of the system.

5. **Operation and maintenance**

   Maintenance is a crucial to ensure that the system remains useful. Maintenance involve correcting errors, which are not discovered in the earlier stage of the life cycle, improving the implementations of the system units and enhancing the system services as a new requirements are discovered.

Other than the overall waterfall model, another waterfall models that show the process of database analysis using data mining techniques is depicted in figure. Notice that this model is a sub-model for the stage of System Analysis and Requirements as the overall waterfall model.

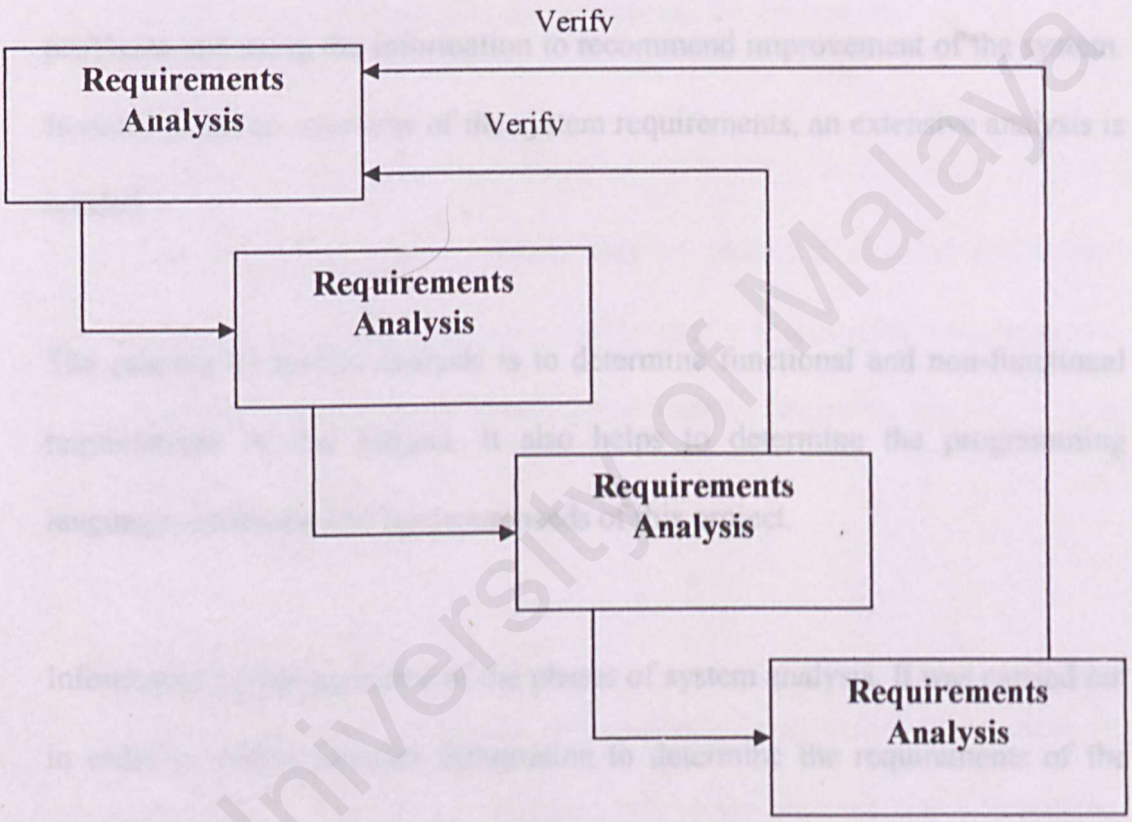


Figure 3.2: Waterfall Model for database analysis using Data Mining Techniques

Chapter 4 – System Analysis and Requirements

4.1 System Analysis

System analysis is the process of gathering and interpreting facts, diagnosing problems and using the information to recommend improvement of the system. In order to get an overview of the system requirements, an extensive analysis is needed.

The purpose of system analysis is to determine functional and non-functional requirements of this project. It also helps to determine the programming language, databases and hardware needs of this project.

Information gathering is one of the phases of system analysis. It was carried out in order to gather relevant information to determine the requirements of the system.

The following is the approach used to define Medical Diagnosis using Data Mining Techniques system requirements: -

- Research

  Research has been done through reviewing books and journals that contain relevant information needed in this system. Besides that, Document Room (FSCIT Library) in FSCIT serves as a good place to do some research from the documentation by the seniors.

- Surfing Internet

  Internet surfing is an efficient way of gathering updated and useful information. Many websites provide useful information and expertise, which is needed in developing the system. Websites that have been visited are listed.

- Informal Interview

  Made an informal interview with expert programmer to know more on data mining tools. Helps to find some useful tools such as Clementine, SQL Server 2000 Analysis Service that can be use as main data mining tools. This helps in understanding the need for the system from user's point of view.

4.2 Requirements Analysis and Specification

A requirement is a feature of the system or a description of something the system is captured of doing in order to fulfill the system's purpose. Based on the gathered information, Medical Diagnosis using Data Mining Techniques requirements have been determined.

There are two ways to describe the requirements, for examples functional and non-functional requirements. The next section describes the functional and non-functional requirements for the proposed system.

4.2.1 Functional Requirements

Functional requirements describe an interaction between the system and its environments. Further more, functional describe how the system showed its behaviors.

This project consists of two major tasks, which is including data mining modeling and system development. The system development is separated to five modules such as Input Module, Selection Module, Training Set Generator

Module, Configuration Module, Data Mining Model Generator Module, and Output Viewer Module. Below are the functional requirements respectively: -

- Input Module

  This module is necessary to let the program engine to identify the flat file regardless of its configuration. This means that a temporary or project database is created according to the project, which the user will be working on.

- Selection Module

  This module is designed to identify the important fields in the project database. This module gets the user to specify the related fields needed either in training mode or in the running mode.

- Training Set Generator Module

  This module generates a data set used to estimate or train a model. Generates a set of tables by using <iis> ...<>. The result table is called training set, which is used for data mining. Notice that converting field to null, discrete and meaningful value generates training set.

- Configuration Module

  In this module, used to set the data mining server name and database name, which will set as global variables.

- Data Mining Model Generator Module

  This module implement the data-mining model chosen to generates the training set. Setting some properties can generate data mining model. For example, data mining algorithms as properties and clustering as value.

- Output Viewer Module

  Let analyzer to view data mining results for client clustering by using simple tables with filter, bar chart, line chart, and pie chart.

4.2.2 Non-functional Requirements

Non-functional requirements are the constraints under which a system must operate and the standards which must be meet by the developed system.

The purposed solutions for this project must have the following non-functional requirements.

i. Accuracy

   Refers to the precision of computation and control. This system should provide various accuracy measures. Data mining tools should also provide accurate prediction and classification results.

ii. Flexibility

This system will be able to incorporate new technologies in the future and in the fast changing environments. This technology includes object-oriented technology and Advanced Security technology.

iii. Usability

The developed system should be user friendly where the user must be able to use the system in the shortest learning courses. The data-mining model can also be customized to meet the need of changing requirements.

iv. Correctness

The system and model must meet the objectives, specification and requirements for user stated earlier.

v. User friendliness

Although this project will not involve o lot of user interaction, it should also apply the Graphical User Interface (GUI) approach for better visual effect to the user. An attractive, simple and easy-to-use interface will be required for the system. Usage of suitable and meaningful captions and icons helps the user to use the system with more confidence.

vi. Modularity

Modularity involves breaking the logical, manageable portions or modules so that distinct functions of objects could be isolated from one another. This characteristic makes testing and maintenance much easier.

4.3 System Development Tools

4.3.1 Clementine

Clementine is a data mining toolkit. It was design to helps to makes the most profitable use of what is, a greatly underutilized resource-data. Clementine's "visual programming" interface is easy to use but supports rich facilities for manipulating data , experimenting with different combinations of data and of techniques, and for testing hypothesis. This facility helps to make it easy to apply knowledge and expertise to discover interesting and valuable properties of your data.

Clementine can also give more active assistance. It includes advanced modeling or "machine learning" technologies, which extract complex interrelationship and decisions-making rules from the data. These help to automate applications

such as prediction, estimation and classification, and can be used to provide "expert" decision support.

Clementine has many facilities for handling such data. Typical things to derive are: -

- Rates of change
- Moving averages
- Sequential ("history") values
- Counts how many times event happens
- Intervals between events
- Alternative states

Clementine is a comprehensive, integrated data mining toolkit, providing facilities for: -

- Accessing data
- Manipulating data
- Visualizing the data using interactive graphical displays
- Modeling aspects of the data using machine learning and statistical regression
- Producing textual displays and reports on the data on model performance
- Exporting models as C code

4.3.2 Visual basic 6.0

Visual Basic is a programming language used to create window-based application. Visual Basic eases the process of designing the user interface. Hundreds of functions and the latest technological advances have been added to the language to make it an industrial-strength development suitable for almost any types of application. For example, Data Mining Tree Helper Object 1.0 and Data Mining Model Browser are two controls that added on the Visual Basic 6.0 after installing SQL Server 2000.

4.4 System Requirements

Bellows are the minimal hardware requirements for developing the systems: -

System requirements

USER

Hardware: Pentium-compatible processor or higher and a monitor with 1024 x 768 resolution or higher (support for 65,536 colors is recommended). A CD ROM drive for installation is also required.

Operating system: Windows 95, Windows 98, Windows 2000 or Windows NT 4.0 with Service Pack 3 or higher.

Minimum free disk space: 80MB

Additional option: Clementine Application Template for Web Mining: 115MB

Additional option: Clementine Application Template for Telecommunications: 60MB

Minimum RAM: 256MB

SERVER

Hardware: Pentium-compatible processor or higher for Windows, SPARC for Solaris, HP Workstation for HP/UX or IBM RS/6000 for AIX. A CD-ROM drive for installation is also required.

Operating system: Windows 2000 or Windows NT 4.0 with Service Pack 6 or higher; Solaris 2.6, 7 or 8; HP/UX 11.0 or 11i; AIX 4

Minimum free drive space: 25MB for installation; plus at least twice the drive space of the amount of data to be processed.

Minimum RAM: 256MB

Chapter 5 – System Design

System design is a very important in system as it determines the success of the system. The system specification describes the features of the system, the component or elements of a system and their appearance to users. Requirements that are found in analysis stage are the one actually translated into design specifications.

This chapter discusses the system functionality design, data flow diagram, and also the database design.

5.1 Data Flow Diagram

A data flow diagram is a graphical technique that depicts information flow and transforms that is applied as data movement from input to output.

Data flow diagram representing a system at any level of detail with a graphic network of symbols showing data flows, data stores, data process, and data sources. The data flow diagram is analogous to a road map. It is a network module of all possibilities with different details shown in different hierarchy levels. The process of representing different level is called "leveling" or "partitioning" by some data flow diagram advocates.

5.1.1 Context Diagram

Context diagram is the highest level in a data flow diagram and contains only one process, representing the entire system. The diagram does not contain any data stores and its fairly simple to create, once the external entities and the data flow to and from then are known.

Below is the context diagram for Medical Diagnosis using Data Mining Techniques: -

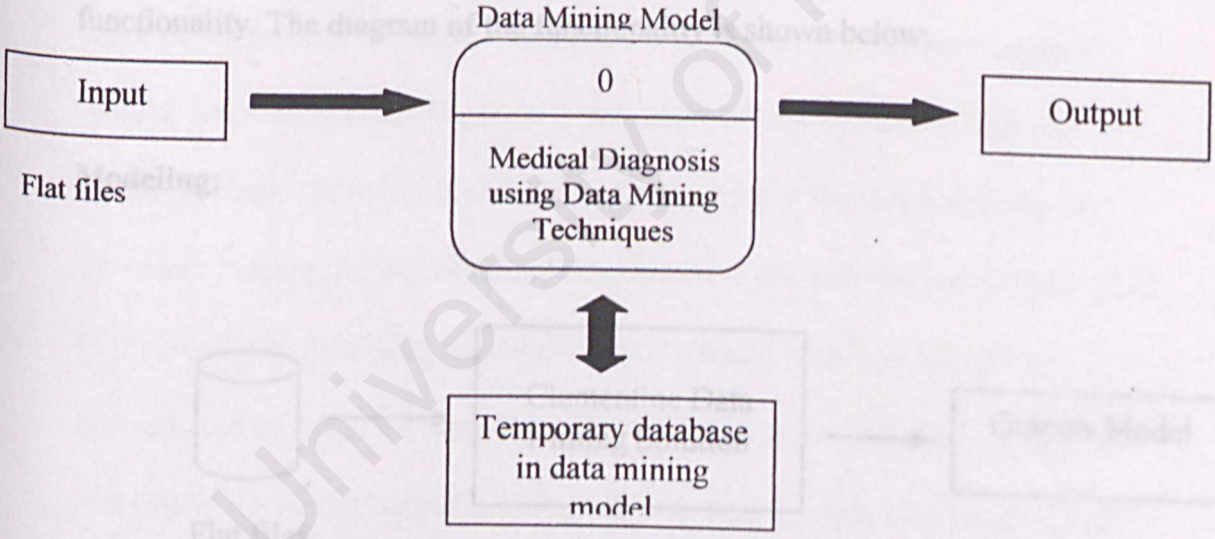


Figure 5.1: Context Diagram

5.2 System Functionality Design

The system functionality design use to identify the major modules involved in developing the system. The major function is the data mining engine in the Clementine Data Mining Solution itself. The components can then be broken or classify into the sub-component that are in turn be broken down more if needed.

Basically the system consists of two different segment of functionality. It is divided into the modeling functionality and generating the user interface functionality. The diagram of the functionality is shown below:

Modeling:

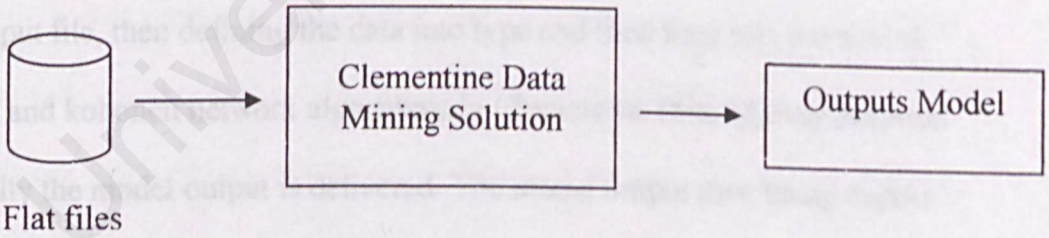


Figure 5.2: System functionality for generating the model

User Interface:

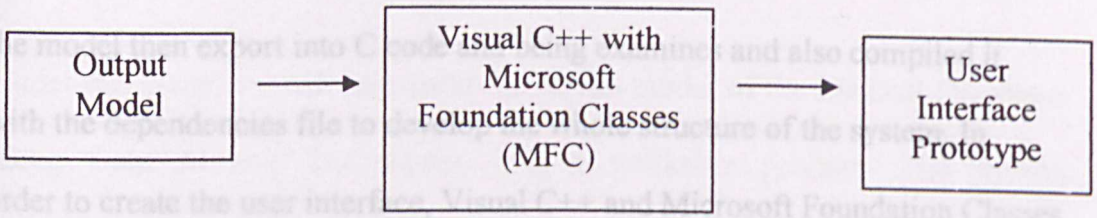


Figure 5.3: System functionality for generating the graphical user interface

5.2.1 Modeling Functionality

As for the modeling functionality, the flat files or the variables files act as the input of the process. Then the stream is created within the Clementine Data Mining Solution in order to generate the output for the system. Within the stream is where the major process take place which is the data being process into more meaningful information. The process is discards the unnecessary field in the input file, then defining the data into type and then feed into the neural network and kohonen network algorithms in Clementine Data Mining Solution and finally the model output is delivered. The model output then being export into C code for creating the user interface.

5.2.2 User Interface Functionality

The whole process starts when the output model is created. From Clementine the model then export into C code and being examines and also compiled it with the dependencies file to develop the whole structure of the system. In order to create the user interface, Visual C++ and Microsoft Foundation Classes (MFC) is used. We created a prototype user interface and integrate with the whole structure of the coding in order to make the system works.

5.3 Database Design

Database is a central source of data meant to be shared by users for variety of applications. Database design involves identifying the user data requirements and determining how these data should be structured. It transformed the unstructured data and the processing requirements of an application. The database model used in this system is a flat file database (text file).

The database will contains the fields of data such as sex, age, blood pressure, cholesterol level, and drug respond.

Chapter 6 – Implementation

Under this stage, we will transform the design model of the Medical Diagnosis using Data Mining Techniques, into a workable product. The system implementation of the Medical Diagnosis using Data Mining Techniques will be divided into two components, which are the platform development and the modules implementation.

6.1 Platform Development

Before the development of the whole application can begin, the environment and the platform for the development phase have to be set up. As the design of the system shows, basically there were two parts that had to be implemented, the browser side and the database.

6.1.1 Operating System Implementation

The first phase to set up the developing phase is to install and set up the Windows 2000 Operating System. This process begins with formatting the hard disk drives of the computer to NT File System (NTFS) to provide more security

feature supported in Windows 2000. It was done by installing through the CD ROM. After the installation process completed, the hardware specification was configured in Windows 2000. Finally the Windows 2000 Service Pack 2 was installed into the computers. After numerous starts, then the whole system was ready.

The development environment for this system resides on an X86-based PC with Windows 2000 as its operating system with 256MB memory. The aforementioned was chosen because X86-based PCs have become the trend in desktop computing. Its popularity is because of its reliability and cost effectiveness compared to other platforms such as Apple Macintosh or other platform. With Windows 2000 as its operating system, stability will not be an issue because it is one of the best operating system around. The overall platform specification is adequate in order to develop a data mining application.

The hardware used was listed here:

Processor	:	Intel Pentium 4 1.6GHz
Hard Disc Storage	:	40 GB IDE Hard Disc
Memory	:	256 DDR SDRAM

Operating System	:	Windows 2000 with Service Pack 2 (SP2)
Application	:	Clementine Data Mining Solution with
		Exceed X Windows Server
		Microsoft Visual Studio v. 6.0

6.1.2 Data Preparation and Implementation

The main database for the system is a very simple set of data stored as a "flat file" or text file. This data set was transform from the data in the databases or data warehouse with all the field was separated with comma delimiter (,). The system also can retrieve data that registered through the ODBC, and also form Microsoft Excel. But for this project, we are using the flat files as the data source.

Upon completion of this project, we are using three kinds of data sets. These data sets are the training sets for training the neural network, and the test set for testing the network performance and reliability. Finally the implementation data sets which the system will use to generate the prediction.

6.1.3 Clementine Data Mining Solution and Exceed

As discussed earlier, the data modeling is done by Clementine, and running Clementine on Windows platform requires an X Windows server. An X Windows server transforms 386, 486, Pentium, or PS/2 computer into an X Window terminal. It provides access to UNIX-based applications (X clients) from within the familiar Microsoft Windows environment. The X Windows server provided with Clementine installation CD is Exceed version 6.0.

6.1.4 Visual Studio – Visual C++ 6.0

Clementine models are exported as C codes and require an ANSI C compiler to compile it. The compiler chosen for this task was Microsoft Visual C++ 6.0. Visual C++ was chosen because it implements object oriented programming methods and offers further enhancement of the system with Microsoft Foundation Class (MFC), which will provide windows-like interface. Although MFC might produce unpredictable results or instability compared to a straightforward implementation using plain dialogs, it offers a wide range of enhancements especially in the graphical user interface (GUI) side.

6.1.5 Modeling and Exporting Clementine Models

The first step to create a model is to specify the training dataset for the model. This is achieved by specifying the training dataset using the 'Variable File' node in Clementine. After specifying the training dataset, the appropriate processing nodes are applied to the training dataset to derive the most suitable pattern. The whole process of discovering the pattern is fully automated by Clementine.

After a suitable model is derived, the model can then be exported into C code so that we can embed the model into our custom-made application. Although the functionality of our custom-made application can be achieved in Clementine alone, it is useful to publish or reuse the model into another stand-alone application without the need to run Clementine.

Models created by Clementine can be accessed from an external program using a few techniques. There are either using the Clementine Solution Publisher or model export into C code. The C code method is chosen because the Clementine Solutions Publisher node in Clementine Data Mining is disabled since the license is expired.

6.2 Standard and Procedure to Write a Code

Standard and procedure will help to organize out thoughts and avoid mistakes or errors. Documenting our code will be some of the procedure so that our code will be clear and easy to follow. Beside that, a standardized documentation will help in locating faults and making changes because it clarifies which functions of our program perform certain functions. With standard and procedures, it will help in translating a design to code. Changes in design are easy to implement in the code because there is a correspondence between design components and the code components.

6.2.1 Coding

The models that generated from Clementine and exported into C code consists of a header file <modulname.h> and a source file <modulname.c>. Depending of the models generated by Clementine, different model type needs different type of file dependencies.

First we have to examine the structure of the model to understand the way to access the model via function call. There were also sample of codes on how to manipulate the models with the samples code in the Clementine User Guide.

Model type	Files generated by Clementine	Other dependencies
C 5.0 Rule	<*modelname*>.c <*modelname*>.h	-
Neural Network	<*modelname*>.c <*modelname*>.h <*modelname*>.san	neural.c neural.h forprop.c forprop.h
Kohonen Network	<*modelname*>.c <*modelname*>.h <*modelname*>.san	koh_net.c koh_net.h kohprop.c kohprop.h

Table 6.1: File model generated from Clementine and its dependencies

Chapter 7– System Testing

System testing is a critical element of software quality assurance was performed to represent the ultimate review of specification, design and coding for the Medical Diagnosis using Data Mining Techniques. The whole system was developed using the Waterfall model in correlation with the data mining process. Modules in the system were tested and enhanced repeatedly until it could function precisely. After all these unit testing, all the modules in the system are then integrated to form an application. The integrated system was then tested in integration testing.

7.1 Unit Testing

The objective of unit testing is to find faults and errors in the Medical Diagnosis using Data Mining Techniques modules. These types of testing method were used as test modules. These methods are source code examining, test casing, and the user testing.

7.1.1 Source Code Examining

The source code examining methods was used as a basic method to test the modules in Medical Diagnosis using Data Mining Techniques application in order to identify faults or inefficiency in the source codes.

The procedures of the source code of each module were compared to the original design of the module process flow to determine the correctness of the source code. Comments were inserted into the sections of the code tested to ensure it could be easily traced in the future.

In C code, the programmer can trace all code, line by line after running the program. Break point also can be inserted to stop the program in particular line. Other functions are also facilitated to allow programmer to view the code.

Visual C++ has provided few debugging tools while examining the code. For example, it allow programmer to examine the code using Debug Object. The Debug Object sends output to the immediate windows at run time.

7.1.2 Test Cases

In order to test the modules in a particular way, test cases were design to test the modules. These test cases were consists of structural data in various range. In this method, the modules were tested with all possibility of input data. Possible situation and data that could cause error or fault could be detected. The test data used also contain data that was normally input by the users. The response of the module to the test data then was tested. Therefore program fault that would happen in normal condition were detected and corrected.

7.1.3 Data Reliability Testing

Basically the system consists of two different structured of system which is the in Visual C <modulenames>.c and Visual C++ <modulesnames>.cpp. The model in Visual C is executed within the DOS mode while the model in Visual C++ is executed with the user interface design. This situation allows us to test the reliability of the whole system in producing the results or output. A set of test data input in both the system and examine the results of both system. From the entire test data tested with both of the system, we figured out that 97% of

the results are reliable. This show that both of the system giving almost the same output even though the structure is different.

7.1.4 User Testing

Test cases are still unable to precisely stimulate a true situation for the Medical Diagnosis using Data Mining Techniques without the involvement of the users. Therefore, it was still inadequate to identify the faults that may occur in all conditions.

With user testing method, user is involved in testing the modules. The users include team members and members of the faculty. Each module was presented to user and allows them to operate these modules by themselves. The capability of detecting error of these modules was then tested. With this, testing effectiveness of these modules in improving user learning curved was also tested.

7.2 Integration Testing

When all the modules were tested and satisfied the requirements, they were integrated into the whole system. During the integration, integration testing was carried out in order to ensure the components of the system would support each other. The objectives of integration testing were as following:-

- Compare the whole system with the functional and non-functional requirements.
- Detect any faults or bugs in the integrated system.
- Exam the correct flow of the integrated system.

The integration testing is also testing the whole structured code in Visual C++ and MFC with the user interface that has been developed. This testing was done to make sure the system does not malfunction when the whole system implemented.

Chapter 8 – System Evaluation

Upon completion of the project, the system strengths and limitation were identified and evaluated. Comparing the final application and the system requirements written earlier identified these strengths and limitation of the system.

8.1 System Strength

The implemented system is an example how neural networks being implemented and applied in solving problems in medical fields. The system helps us automate and understand more meaningful data in different pattern that hidden behind set of data. This will helps in giving pattern such as the prediction techniques which is very useful in medical fields.

The main feature that is very important is the model generated from Clementine Data Mining Solution. Although the model being export in C codes and become static, still it can give higher level of accuracy and simple. The model generated from Clementine Data Mining is independent and does not need the user to

install it. The model can be distributed to several workstation and desktop computer with only one workstation

The user interface of the system let the user act as minimum as possible to view the information required. In order to increase the efficiency of the information produced by the system, the helps of the medical expert is still needed in interpreting the pattern in making the decision.

8.2 System Limitation

Even though the strength of the system depends on the exported model but still it can be outdated from time to time. This means that the model created is static and not be able to learn more patterns using the neural network algorithms in the Clementine Data Mining Solution. In order to prevent the model from outdated, the dependency on Clementine is still needed in producing the updated model from time to time as the algorithms learn the pattern of data. In worst cases, a new model should generate from the new stream thus creating possibilities in developing the new application structure for the model.

In order to make the prediction more accurate, the system needs to specified details in each of data sets. As an example the data sets can not be mixed up with several different drugs for several different symptoms. Therefore, we may need different approach of system due to various kind of symptoms occurs.

8.3 Future Enhancement

In order to improve upon the current developed system and the project, few enhancement or new added features could be possibly implemented. The following are some enhancement proposed:-

- Since Clementine Data Mining solution is licensed software, in future may be there an open source might be appropriate in replacing Clementine for modeling the streams.
- Can manage pragmatically by using data mining model prediction and clustering.
- More samples files or data sources supported by the data mining process to enhance the functionality.

8.4 Problem Encountered

Along the process of completing this project, a lot of problems surfaced from time to time. These problems usually involved the use of the new technologies and products such as the Windows 2000, Windows 2000 Server and others.

Another problem is the stability of the products. Some of these products were not stable; sometimes they intend to malfunction. However, some of these problems still remain unsolved. The following list of some of the problems faced:-

- In the process of installing the system tool, some unexpected failures were encountered. The installation and setup process had to be repeated in order to solve the problems.
- There is no company willing to offers to share out their documents for the project as the project involves the medical data for the data source. The reason is the patients rights and the policies that the privacy of the data to public.

- There is not enough time for studying and implemented the system at the same time. We also encountered problem in choosing the implementation tools.
- Tool that is chosen for data mining is quite new and there are insufficient references about the tool for reviewing the project.
- Although we form a group work for the project, but still not enough time and resources to be discussed for our project.
- Time of completion of the system is limited due to the semester schedule. The implementation and the testing phase are too short.

8.5 Objectives Achieved

In general, the system was developed in order to help the medical expert and management level in discovering the new pattern and analyst the prediction pattern for the use of giving drugs prescription for the future patient.

- Discover and maintain the prediction pattern of the drugs taken by the patient based on several factor such as age, sex, blood pressure, cholesterol level, and others.

- The use of data mining techniques in optimizing the pattern even though the generated model from Clementine is a static model.
- Survey on available and useful techniques for such tasks.
- Utilize the result of data mining. Created an analysis module based on the data source pattern resulted form the data mining techniques.

Chapter 9 – Conclusion

Finally, the project Medical Diagnosis using Data Mining Techniques has been implemented and completed. Even though, it is still not a complete solution for managing or predicting the best solution, but with further enhancement, it can be a more powerful tool to enable this function.

During the completion of the project, a lot of opportunity was given to learn new technologies such as the Clementine Data Mining Solution, Weka 3-2 in appreciation with Java Programming tools, Visual C++ and others. I learned a lot from these hand on experience, even more than what I have learned from class lectures.

A lot of time spent to figure out how it works and to actually configure it. More time was then spent to troubleshoot the problems faced.

From this project Medical Diagnosis using Data Mining Techniques the most utmost useful, I believe, it's guiding me handling major tasks in future. One of the most essential knowledge and experience gained during this project is to

learn the way of handling a project and apply the theories learned from the courses.

In my opinion, as undergraduates, should appreciate this learning process which gives us the opportunity to develop and improve our skills in designing, implementing, and operating a project successfully under the guidance and supervision of the lecturer. Moreover, thesis also serves as a channel for students to apply both theoretical skills learned into a time-budgeted project.

Throughout this project, knowledge and experience have been acquired. The benefits are listed:-

- Opportunity to learn additional programming languages and other tools.
- Hand-on experience in planning and handling a project.
- Improving the time management of certain tasks.
- Chances to know and use new technologies and several different software, which are not familiar before.
- Experiencing the real world problem solving concepts.
- Skilled for documentation in writing and presentation of project.

References

Book:

Michael J.A. Berry and Gordon S. Linoff, *Mastering Data Mining: The Art and Science of Customer Relationship Management.* John Wiley & Sons Inc.,2000

http://www-db.stanford.edu/~ullman/mining/cluster1.pdf

Shaari Lawrence Pfleeger, *Software Engineering : Theory and Practice.* Prentice Hall, 2001.

Sommerville, I. *Software Engineering 5th Edition,* Addison-Wesley Ltd.

E.Kendall and J. E. Kendall, *System Analysis and Design, Third Edition.* London, Prentice Hall, 1995.

Jeffry L. Whitten, Lonnie D. Bentley and Kevin C. Dittman, *System Analysis and Design Methods, 5Th Edition.* Mc Graw-Hill, 2001.

http://www.mhhe.com/whitten

Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition,* The Two Crows Corporation, 1999.

Ivor Horton, *Beginning C, 2nd Edition.* Wrox Press Ltd,1997

WWW:

http://www.microsoft.com/sql

http://www.spss.com/datamine/techniques.htm

http://www.statofinc.com/textbook/stdadmin.html

http://www-4.ibm.com/software/data/iminer

http://www.twocrows.com

http://www.ncdm.uic.edu/dataminingresearch.htm

http://www-db.stanford.edu/~ullman/mining/cluster1.pdf

http://www.google/product/datamining

Appendix:

Appendix A: Data Sets

Appendix B: System Screen Shots

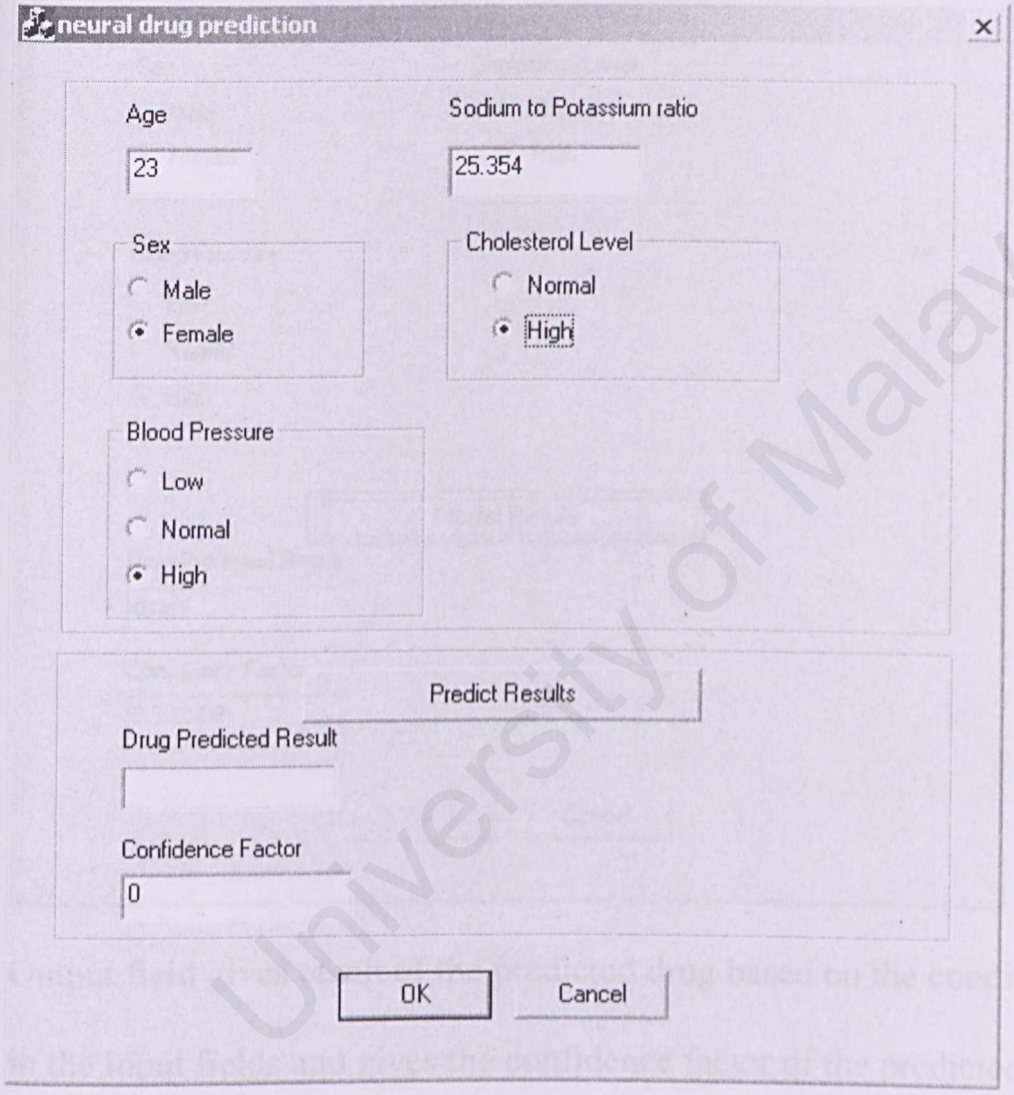
```
Age,Sex,BP,Cholesterol,Na,K,Drug,Na_to_K
23,F,HIGH,HIGH,0.792535,0.031258,drugY,25.354629
47,M,LOW,HIGH,0.739309,0.056468,drugC,13.09253
47,M,LOW,HIGH,0.697269,0.068944,drugC,10.113556
28,F,NORMAL,HIGH,0.563682,0.072289,drugX,7.797618
61,F,LOW,HIGH,0.559294,0.030998,drugY,18.042906
22,F,NORMAL,HIGH,0.676901,0.078647,drugX,8.606825
49,F,NORMAL,HIGH,0.789637,0.048518,drugY,16.275135
41,M,LOW,HIGH,0.766635,0.069461,drugC,11.036913
60,M,NORMAL,HIGH,0.777205,0.05123,drugY,15.170896
43,M,LOW,NORMAL,0.526102,0.027164,drugY,19.367619
47,F,LOW,HIGH,0.896056,0.076147,drugC,11.76745
34,F,HIGH,NORMAL,0.667775,0.034782,drugY,19.198867
43,M,LOW,HIGH,0.626527,0.040746,drugY,15.376405
74,F,LOW,HIGH,0.792674,0.037851,drugY,20.941957
50,F,NORMAL,HIGH,0.82778,0.065166,drugX,12.702636
16,F,HIGH,NORMAL,0.833837,0.053742,drugY,15.515556
69,M,LOW,NORMAL,0.848948,0.074111,drugX,11.455088
43,M,HIGH,HIGH,0.656371,0.046979,drugA,13.971583
23,M,LOW,HIGH,0.55906,0.076609,drugC,7.297576
32,F,HIGH,NORMAL,0.643455,0.024773,drugY,25.974044
57,M,LOW,NORMAL,0.536746,0.028061,drugY,19.127829
63,M,NORMAL,HIGH,0.616117,0.023773,drugY,25.91667
47,M,LOW,NORMAL,0.809199,0.026472,drugY,30.56811
48,F,LOW,HIGH,0.87444,0.058155,drugY,15.036368
33,F,LOW,HIGH,0.858387,0.025634,drugY,33.486268
28,F,HIGH,NORMAL,0.556833,0.029604,drugY,18.809384
31,M,HIGH,HIGH,0.740936,0.0244,drugY,30.36623
49,F,NORMAL,NORMAL,0.694689,0.074055,drugX,9.380717
39,F,LOW,NORMAL,0.649096,0.028598,drugY,22.697252
45,M,LOW,HIGH,0.753504,0.041976,drugY,17.950829
18,F,NORMAL,NORMAL,0.553567,0.063265,drugX,8.749972
74,M,HIGH,HIGH,0.715337,0.074773,drugB,9.566782
49,M,LOW,NORMAL,0.625889,0.056828,drugX,11.013743
65,F,HIGH,NORMAL,0.828898,0.026004,drugY,31.875788
53,M,NORMAL,HIGH,0.644936,0.045632,drugX,14.133415
46,M,NORMAL,NORMAL,0.526226,0.072234,drugX,7.285018
32,M,HIGH,NORMAL,0.52975,0.056087,drugA,9.445148
39,M,LOW,NORMAL,0.604973,0.043404,drugX,13.938185
39,F,NORMAL,NORMAL,0.517515,0.053301,drugX,9.709293
15,M,NORMAL,HIGH,0.64236,0.07071,drugX,9.084429
73,F,NORMAL,HIGH,0.832683,0.043321,drugY,19.221232
58,F,HIGH,NORMAL,0.868924,0.061023,drugB,14.239287
50,M,NORMAL,NORMAL,0.747815,0.04736,drugY,15.790013
23,M,NORMAL,HIGH,0.593596,0.048417,drugX,12.260074
50,F,NORMAL,NORMAL,0.601915,0.048957,drugX,12.294769
66,F,NORMAL,NORMAL,0.611333,0.075412,drugX,8.106575
37,F,HIGH,HIGH,0.559171,0.042713,drugA,13.091354
68,M,LOW,HIGH,0.726677,0.070616,drugC,10.290543
23,M,NORMAL,HIGH,0.888629,0.028045,drugY,31.685826
28,F,LOW,HIGH,0.606933,0.030659,drugY,19.796243
58,F,HIGH,HIGH,0.560854,0.028886,drugY,19.416119
67,M,NORMAL,NORMAL,0.846892,0.077711,drugX,10.897968
62,M,LOW,NORMAL,0.804173,0.029584,drugY,27.1827
24,F,HIGH,NORMAL,0.648646,0.035144,drugY,18.456806
68,F,HIGH,NORMAL,0.77541,0.0761,drugB,10.189356
26,F,LOW,HIGH,0.578002,0.040819,drugC,14.160122
65,M,HIGH,NORMAL,0.635551,0.056043,drugB,11.340417
40,M,HIGH,HIGH,0.557133,0.020022,drugY,27.826041
60,M,NORMAL,NORMAL,0.645515,0.063971,drugX,10.090744
34,M,HIGH,HIGH,0.888144,0.047486,drugY,18.703281
38,F,LOW,NORMAL,0.598753,0.020042,drugY,29.874913
24,M,HIGH,NORMAL,0.613261,0.064726,drugA,9.474724
67,M,LOW,NORMAL,0.820638,0.039657,drugY,20.693396
45,M,LOW,NORMAL,0.532632,0.063636,drugX,8.369979
60,F,HIGH,HIGH,0.800607,0.060181,drugB,13.303318
68,F,NORMAL,NORMAL,0.821584,0.030373,drugY,27.049814
29,M,HIGH,HIGH,0.625272,0.048637,drugA,12.855892
17,M,NORMAL,NORMAL,0.722286,0.06668,drugX,10.832124
54,M,NORMAL,HIGH,0.504995,0.02048,drugY,24.657959
18,F,HIGH,NORMAL,0.564811,0.023266,drugY,24.27624
70,M,HIGH,HIGH,0.658606,0.047153,drugB,13.967425
28,F,NORMAL,HIGH,0.860775,0.04375,drugY,19.674857
24,F,NORMAL,HIGH,0.80554,0.07596,drugX,10.604792
41,F,NORMAL,NORMAL,0.844196,0.036857,drugY,22.904631
31,M,HIGH,NORMAL,0.88624,0.051922,drugY,17.06868
26,M,LOW,NORMAL,0.790664,0.037815,drugY,20.90874
36,F,HIGH,HIGH,0.734119,0.065556,drugA,11.19835
26,F,HIGH,NORMAL,0.823793,0.042994,drugY,19.16065
19,F,HIGH,HIGH,0.516973,0.038832,drugA,13.313067
32,F,LOW,NORMAL,0.724422,0.066829,drugX,10.839935
60,M,HIGH,HIGH,0.805651,0.057821,drugB,13.933536
64,M,NORMAL,HIGH,0.5126,0.066049,drugX,7.760905
32,F,LOW,HIGH,0.730854,0.075256,drugC,9.711571
38,F,HIGH,NORMAL,0.733842,0.064793,drugA,11.325946
47,F,LOW,HIGH,0.539774,0.05362,drugC,10.066654
59,M,HIGH,HIGH,0.816356,0.058583,drugB,13.935032
51,F,NORMAL,HIGH,0.678646,0.04991,drugX,13.597395
69,M,LOW,HIGH,0.854733,0.055221,drugY,15.478405
37,F,HIGH,NORMAL,0.795312,0.034443,drugY,23.090672
50,F,NORMAL,NORMAL,0.73961,0.042972,drugY,17.21144
62,M,NORMAL,HIGH,0.755873,0.045551,drugY,16.593994
41,M,HIGH,NORMAL,0.658397,0.043442,drugY,15.155771
29,F,HIGH,HIGH,0.857934,0.029132,drugY,29.449883
42,F,LOW,NORMAL,0.763404,0.026081,drugY,29.270503
56,M,LOW,HIGH,0.812663,0.054123,drugY,15.015114
36,M,LOW,NORMAL,0.52765,0.046188,drugX,11.423963
58,F,LOW,HIGH,0.886865,0.023188,drugY,38.246722
56,F,HIGH,HIGH,0.750962,0.029571,drugY,25.395218
20,M,HIGH,NORMAL,0.764067,0.021439,drugY,35.639116
15,F,HIGH,NORMAL,0.697052,0.041677,drugY,16.7251
31,M,HIGH,NORMAL,0.695183,0.058559,drugA,11.871497
45,F,HIGH,HIGH,0.547821,0.042619,drugA,12.853915
28,F,LOW,HIGH,0.656292,0.049997,drugC,13.126628
56,M,NORMAL,HIGH,0.627866,0.070026,drugX,8.966184
22,M,HIGH,NORMAL,0.860621,0.030417,drugY,28.294079
37,M,LOW,NORMAL,0.616692,0.068765,drugX,8.968109
22,M,NORMAL,HIGH,0.536324,0.044871,drugX,11.952575
42,M,LOW,HIGH,0.756097,0.03778,drugY,20.013155
72,M,HIGH,NORMAL,0.72142,0.074552,drugB,9.676736
23,M,NORMAL,HIGH,0.543355,0.032247,drugY,16.849784
50,M,HIGH,HIGH,0.518285,0.069193,drugA,7.490425
47,F,NORMAL,NORMAL,0.526835,0.078828,drugX,6.683349
35,M,LOW,NORMAL,0.685143,0.074717,drugX,9.169841
65,F,LOW,NORMAL,0.760221,0.055214,drugX,13.768628
20,F,NORMAL,NORMAL,0.581591,0.062667,drugX,9.280658
51,M,HIGH,HIGH,0.83849,0.045831,drugY,18.295259
67,M,NORMAL,NORMAL,0.721257,0.075808,drugX,9.51426
40,F,NORMAL,HIGH,0.512517,0.05073,drugX,10.102839
32,F,HIGH,NORMAL,0.724375,0.070383,drugA,10.291903
61,F,HIGH,HIGH,0.63126,0.02478,drugY,25.474576
28,M,NORMAL,HIGH,0.584179,0.021585,drugY,27.064119
```

15,M,HIGH,NORMAL,0.58301,0.033885,drugY,17.205548
34,M,NORMAL,HIGH,0.602557,0.026833,drugY,22.455819
36,F,NORMAL,HIGH,0.563217,0.033618,drugY,16.753436
53,F,HIGH,NORMAL,0.760809,0.060889,drugB,12.495016
19,F,HIGH,NORMAL,0.742092,0.028576,drugY,25.969065
66,M,HIGH,HIGH,0.84985,0.051988,drugY,16.347042
35,M,NORMAL,NORMAL,0.523623,0.066745,drugX,7.845127
47,M,LOW,NORMAL,0.84773,0.025274,drugY,33.541584
32,F,NORMAL,HIGH,0.549375,0.073474,drugX,7.477135
70,F,NORMAL,HIGH,0.725424,0.035406,drugY,20.488731
52,M,LOW,NORMAL,0.663146,0.020143,drugY,32.921908
49,M,LOW,NORMAL,0.510473,0.037539,drugX,13.598471
24,M,NORMAL,HIGH,0.854591,0.033142,drugY,25.78574
42,F,HIGH,HIGH,0.533228,0.025348,drugY,21.036295
74,M,LOW,NORMAL,0.787812,0.065984,drugX,11.93944
55,F,HIGH,HIGH,0.637231,0.058054,drugB,10.976522
35,F,HIGH,HIGH,0.869854,0.06746,drugA,12.894367
51,M,HIGH,NORMAL,0.832467,0.073392,drugB,11.342749
69,F,NORMAL,HIGH,0.773798,0.076882,drugX,10.064749
49,M,HIGH,NORMAL,0.500169,0.079788,drugA,6.268725
64,F,LOW,NORMAL,0.554182,0.021529,drugY,25.741186
60,M,HIGH,NORMAL,0.635762,0.073744,drugB,8.621203
74,M,HIGH,NORMAL,0.818999,0.053057,drugY,15.43621
39,M,HIGH,HIGH,0.731091,0.075652,drugA,9.663869
61,M,NORMAL,HIGH,0.745123,0.078906,drugX,9.443173
37,F,LOW,NORMAL,0.804155,0.066981,drugX,12.005718
26,F,HIGH,NORMAL,0.781928,0.063535,drugA,12.307043
61,F,LOW,NORMAL,0.522891,0.071238,drugX,7.340057
22,M,LOW,HIGH,0.526672,0.064617,drugC,8.150672
49,M,HIGH,NORMAL,0.538183,0.061859,drugA,8.700157
68,M,HIGH,HIGH,0.639888,0.058123,drugB,11.009205
55,M,NORMAL,NORMAL,0.509181,0.070126,drugX,7.260945
72,F,LOW,NORMAL,0.7586,0.05181,drugX,14.641961
37,M,LOW,NORMAL,0.73154,0.043743,drugY,16.72359
49,M,LOW,HIGH,0.655222,0.062181,drugC,10.537335
31,M,HIGH,NORMAL,0.749717,0.06678,drugA,11.22667
53,M,LOW,HIGH,0.618603,0.026939,drugY,22.963102
59,F,LOW,HIGH,0.640455,0.06132,drugC,10.444472
34,F,LOW,NORMAL,0.825542,0.063881,drugX,12.923123
30,F,NORMAL,HIGH,0.501956,0.048067,drugX,10.44284
57,F,HIGH,NORMAL,0.754166,0.075832,drugB,9.945221
43,M,NORMAL,NORMAL,0.538856,0.041905,drugX,12.858991
21,F,HIGH,NORMAL,0.745098,0.026023,drugY,28.632287
16,M,HIGH,NORMAL,0.561019,0.029516,drugY,19.007284
38,M,LOW,HIGH,0.851019,0.046516,drugY,18.295189
58,F,LOW,HIGH,0.887928,0.033324,drugY,26.645301
57,F,NORMAL,HIGH,0.596099,0.041931,drugX,14.216189
51,F,LOW,NORMAL,0.876828,0.038118,drugY,23.002991
20,F,HIGH,HIGH,0.887426,0.078798,drugA,11.262037
28,F,NORMAL,HIGH,0.744956,0.057843,drugX,12.878931
45,M,LOW,NORMAL,0.71486,0.071367,drugX,10.016674
39,F,NORMAL,NORMAL,0.809196,0.046978,drugY,17.224999
41,F,LOW,NORMAL,0.749905,0.040018,drugY,18.739192
42,M,HIGH,NORMAL,0.85794,0.067203,drugA,12.766394
73,F,HIGH,HIGH,0.808019,0.044038,drugY,18.348222
48,M,HIGH,NORMAL,0.769197,0.073633,drugA,10.446362
25,M,NORMAL,HIGH,0.775702,0.040803,drugY,19.010906
33,M,NORMAL,HIGH,0.609566,0.038171,drugY,15.969348
67,F,NORMAL,HIGH,0.785251,0.049416,drugY,15.890622
22,F,HIGH,NORMAL,0.817625,0.035832,drugY,22.818291
59,F,NORMAL,HIGH,0.882486,0.063563,drugX,13.883643
20,F,LOW,NORMAL,0.811023,0.069402,drugX,11.685874
36,F,HIGH,NORMAL,0.575058,0.037124,drugY,15.490195
18,F,HIGH,HIGH,0.88515,0.023802,drugY,37.188051
57,F,NORMAL,NORMAL,0.551967,0.021317,drugY,25.893278
70,M,HIGH,HIGH,0.589493,0.059854,drugB,9.848849
47,M,HIGH,HIGH,0.56332,0.054152,drugA,10.402571
65,M,HIGH,NORMAL,0.8645,0.024702,drugY,34.997166
64,M,HIGH,NORMAL,0.739914,0.035349,drugY,20.931681
58,M,HIGH,HIGH,0.76909,0.040497,drugY,18.991283
23,M,HIGH,HIGH,0.53406,0.066666,drugA,8.01098
72,M,LOW,HIGH,0.547347,0.03356,drugY,16.309505
72,M,LOW,HIGH,0.50482,0.074573,drugC,6.769474
46,F,HIGH,HIGH,0.773569,0.022302,drugY,34.686082
56,F,LOW,HIGH,0.848774,0.07338,drugC,11.56683
16,M,LOW,HIGH,0.743021,0.061886,drugC,12.006286
52,M,NORMAL,HIGH,0.549945,0.055581,drugX,9.894478
23,M,NORMAL,NORMAL,0.78452,0.055959,drugX,14.01955
40,F,LOW,NORMAL,0.683503,0.060226,drugX,11.348969

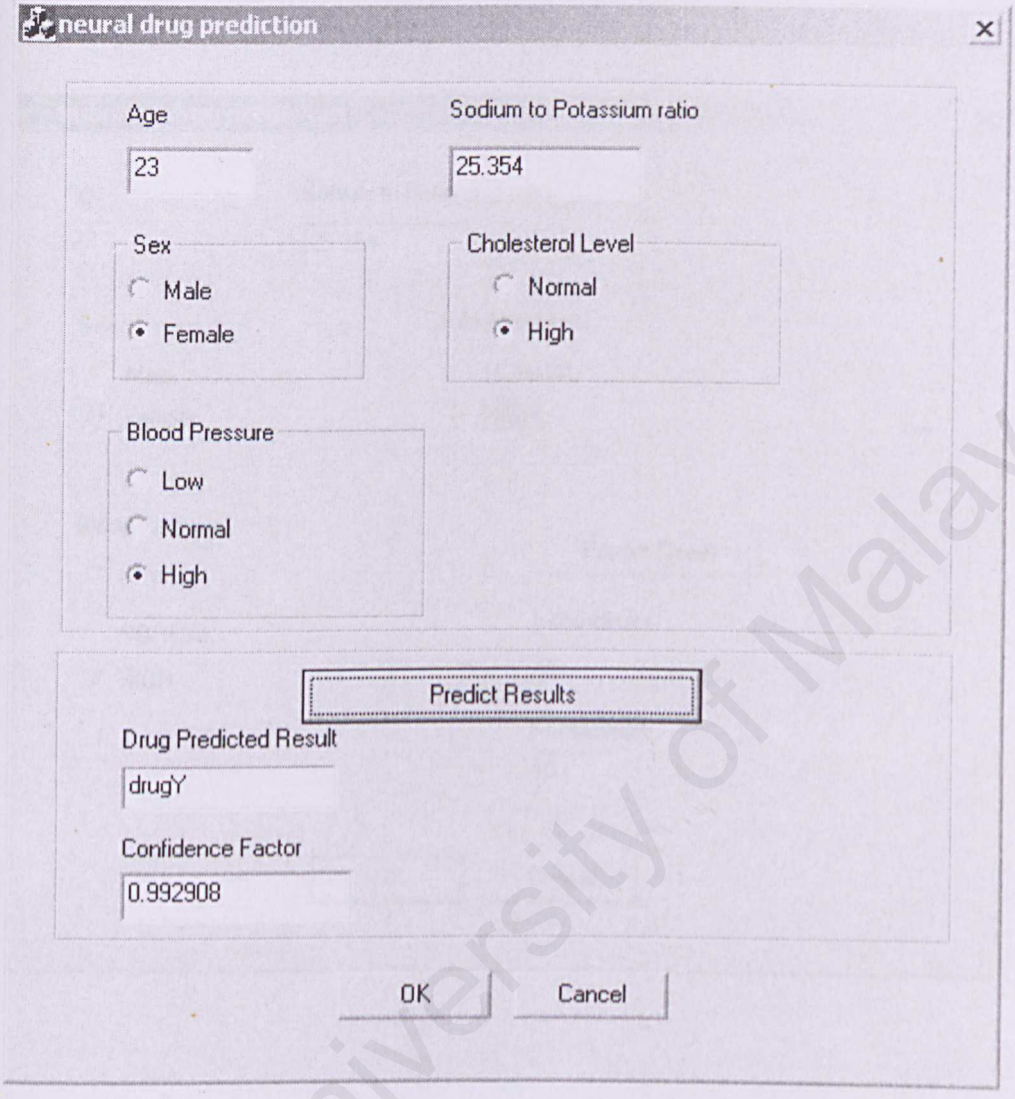
System Screen Shots:

Neural Drug Prediction

Input fields:



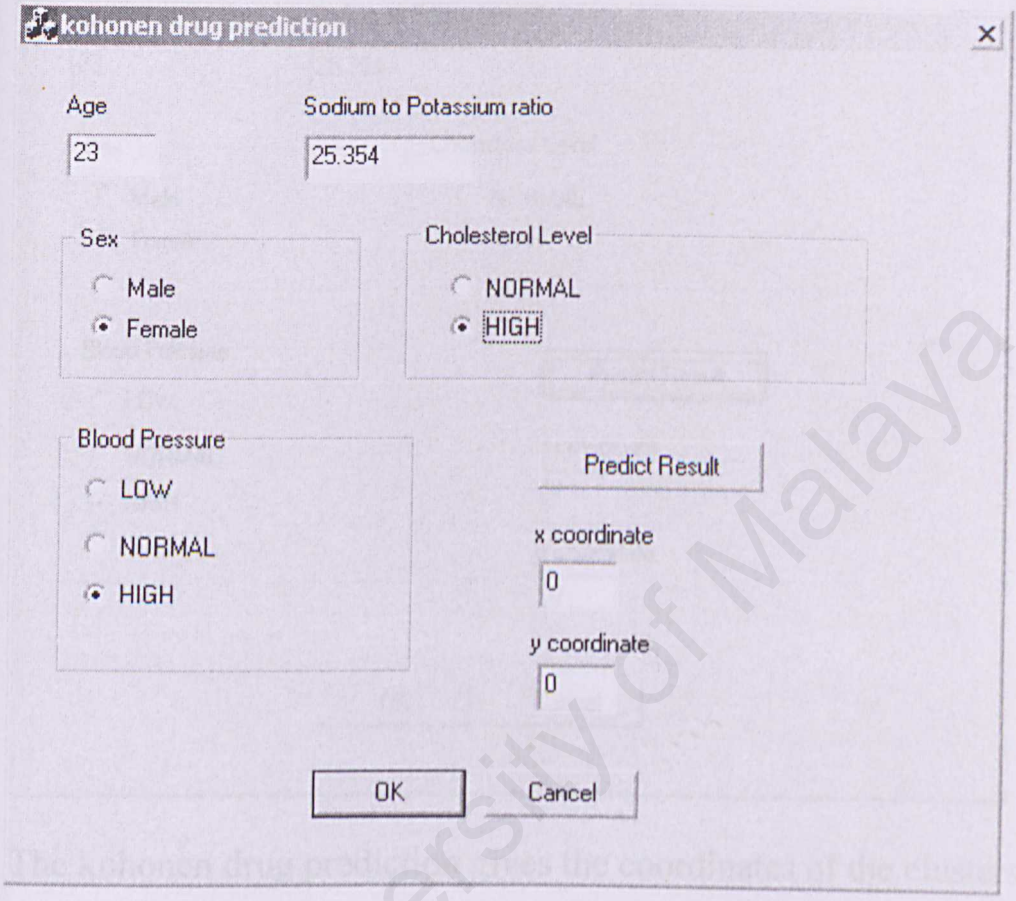
Output fields:



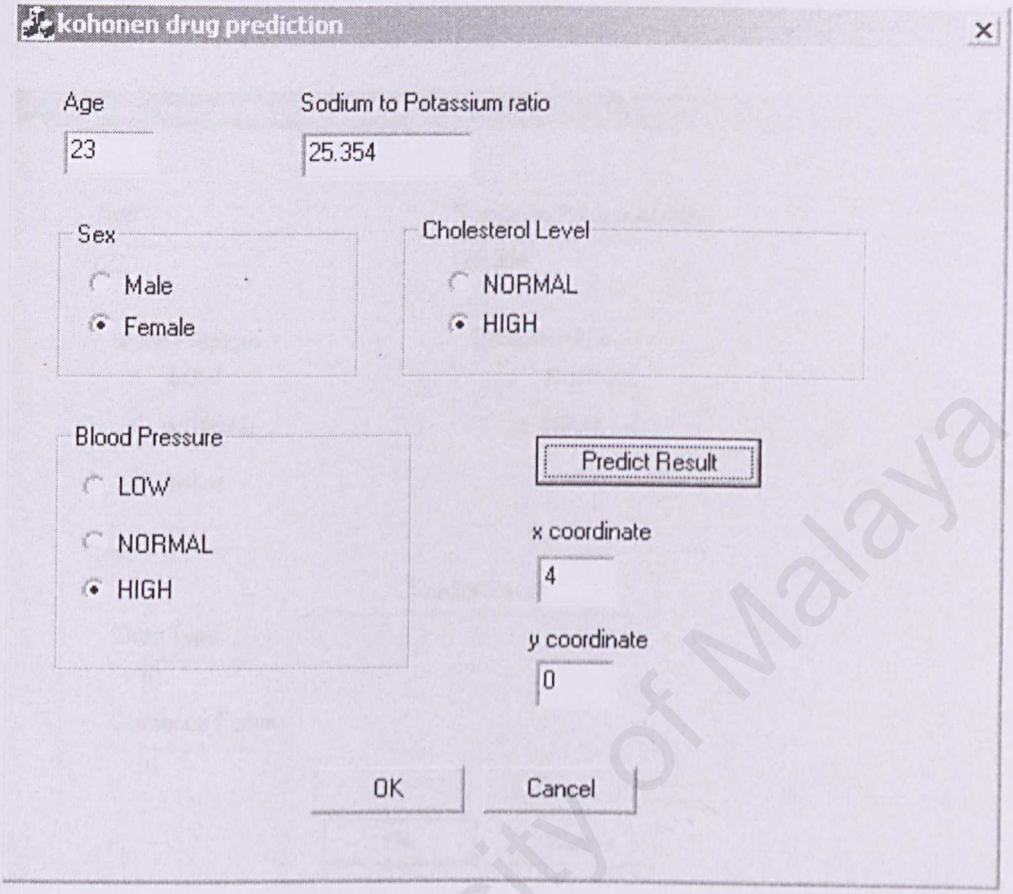
Output field gives result of the predicted drug based on the condition input in the input fields and gives the confidence factor of the predicted result.

Kohonen Drug Prediction (Cluster)

Input fields:



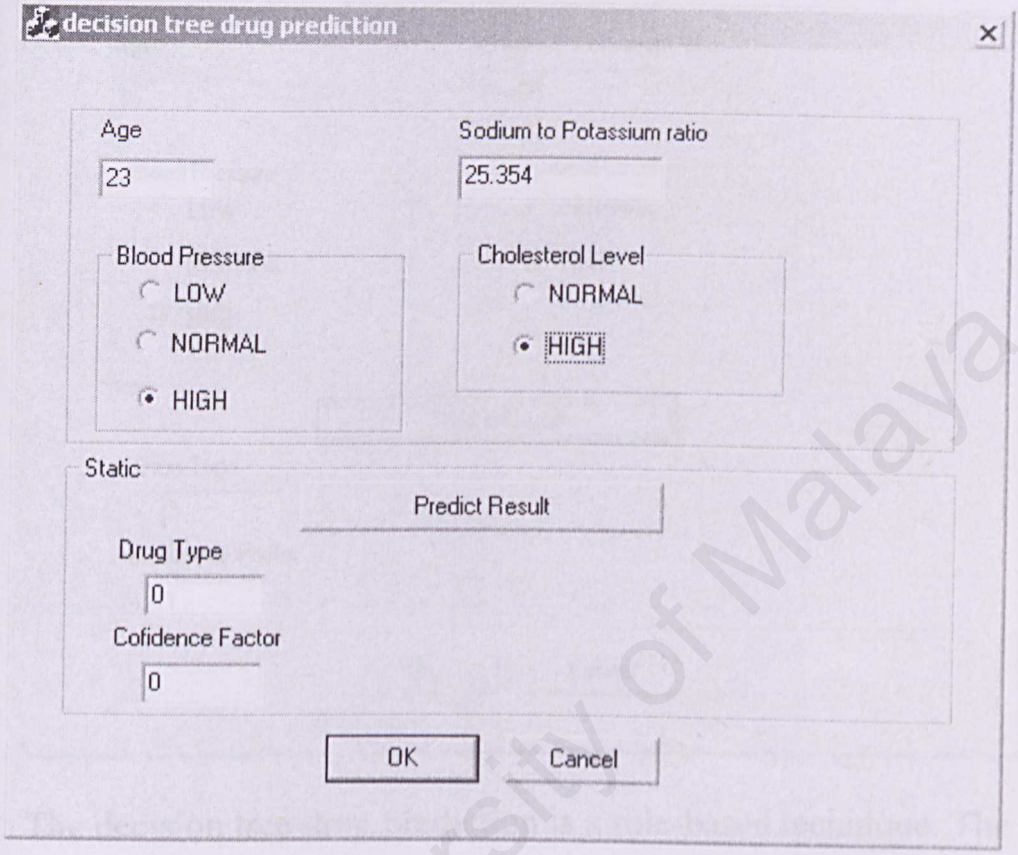
Output fields:



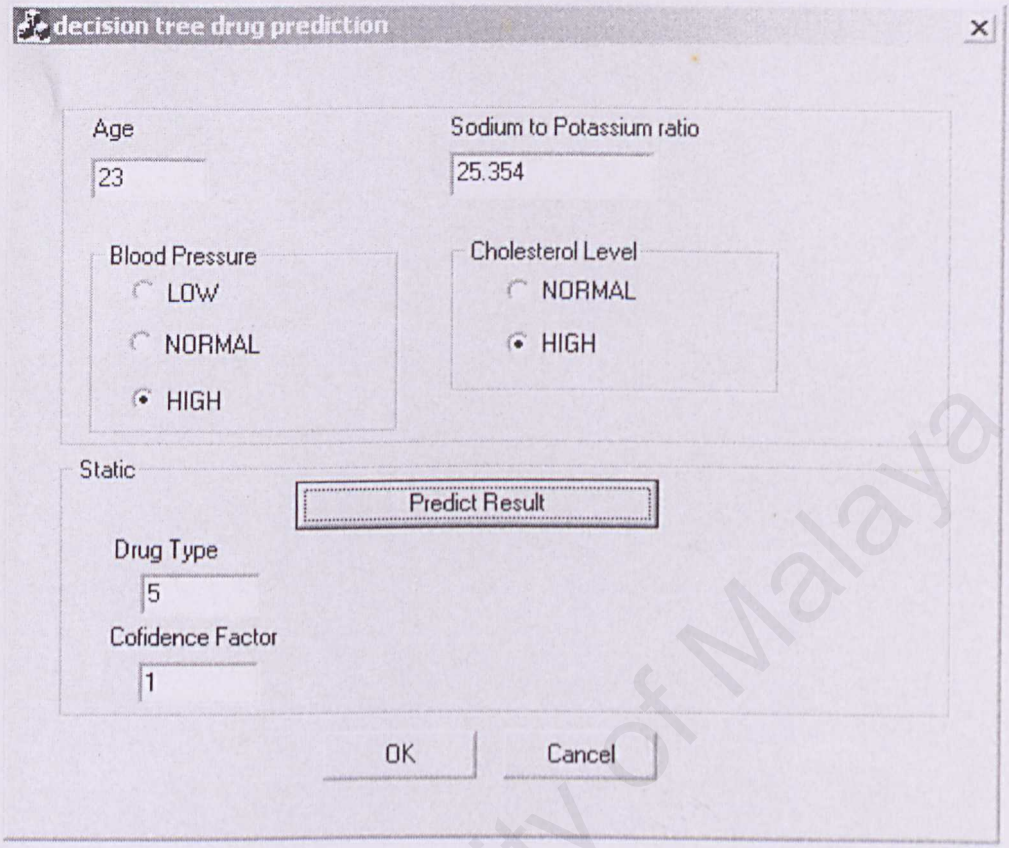
The kohonen drug prediction gives the coordinates of the clusters drug in *k* maps.

Decision Tree Drug Prediction

Input fields:



Output fields:



The decision tree drug prediction is a rule-based technique. The integer indicates in the Drug Type field represent Drug Y. The type of drug represent by the integer is as follows:

0 : UNKNOWN
1 : DRUG A
2 : DRUG B
3 : DRUG C
4 : DRUG X
5 : DRUG Y

The confidence factor given is in ratio 0 and 1.