

**STRENGTHENING MALARIA DIAGNOSIS AND
TREATMENT USING CART**

MYA MYINTZU

M.C.SC.

SEPTEMBER 2022

**STRENGTHENING MALARIA DIAGNOSIS AND
TREATMENT USING CART**

BY

MYA MYINTZU

B.C.Sc (Hons:)

**A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of**

Master of Computer Science

(M.C.Sc.)

University of Computer Studies, Yangon

SEPTEMBER 2022

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis. To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to express my deepest gratitude and my thanks to **Dr. Mie Mie Khin**, Rector, University of Computer Studies, Yangon, for her kind permission to submit this thesis.

I would like to express my appreciation to **Dr. Si Si Mar Win**, Professor and **Dr. Tin Zar Thaw**, Professor, Faculty of Computer Science, University of Computer Studies, Yangon, for their superior suggestions, administrative supports and encouragement during my academic study.

My thanks and regards go to my supervisor, **Dr. Yu Mon Zaw**, Lecturer, Faculty of Computer Science, University of Computer Studies, Yangon, for her support, guidance, supervision, patience and encouragement during the period of study towards completion of this thesis.

I would like to thank, **Daw Hnin Yee Aung**, Lecture, Department of English for editing my thesis paper from language point of view.

I would like to express my appreciation to all faculty members, lecturers, librarians and staff member who have wholeheartedly made their greatest efforts and support during the whole preparation processes.

Finally, I sincerely thank to my beloved husband, **Min Maung Maung Myo** who always support in the moments when there was no one to answer my queries and kept high motivation through this work.

ABSTRACT

Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected mosquitoes. It remains a leading cause of morbidity and mortality in Myanmar. Myanmar is moving towards a more targeted approach to malaria control and is considering malaria elimination as a possible medium-term objective, malaria surveillance becomes central to the program strategy. National Malaria Control Programme and its implementing partners have been participating in strengthening malaria surveillance system in Myanmar in accordance with technical aspects provided in the implementation plan. The most important malaria activity is the early diagnosis and prompt treatment of the disease. Malaria diagnosis and treatment system helps to identify the symptoms of each patient and treatment given. To strengthen Malaria diagnosis, Classification and Regression Tree (CART) algorithm is applied and monitoring treatment processes is carried out with Rule-based algorithm. In this system, annual dataset of malaria patients from Paletwa Township, Chin State of 2017 are used for diagnosis of malaria. Classification with CART algorithm makes the accurate diagnosis and better to follow up.

CONTENTS

	Page
ACKNOWLEDGEMENTS	i
ABSTRACT	ii
CONTENTS	iii
LIST OF FIGURES	v
LIST OF TABLES	vi
LIST OF EQUATIONS	vii
CHAPTER 1 INTRODUCTION	1
1.1 Introduction	1
1.2 Objectives of the Thesis	2
1.3 Method of Study	2
1.4 Organization of Thesis	3
CHAPTER 2 CLASSIFICATION IN DATA MINING	4
2.1 Data Mining Process	4
2.2 Components of Data Mining System	5
2.3 Process of Classification	6
2.4 Method of Classification	9
2.4.1 Classification with Decision Tree Induction	9
2.4.2 Classification with Bayesian	12
2.4.3 Classification with Rule-based	13
2.4.4 Classification with Backpropagation	14
2.4.5 Classification with Support Vector Machine	14
2.4.6 Classification with Association	15
2.4.7 Classification with K-Nearest-Neighbor	16
2.4.8 Classification with Case-Based Reasoning	17
2.5 Other Classification Method	18

2.5.1	Classification with Genetic Algorithms	18
2.5.2	Classification with Rough Set Approach	19
2.5.3	Classification with Fuzzy Set Approach	19
2.6	Review on Data Mining in Healthcare	20
CHAPTER 3	DECISION TREE USING CART ALGORITHM	22
3.1	Decision Tree Learning	22
3.2	Classification and Regression Tree (CART) Algorithm	23
3.3	Types of CART Variables	25
3.3.1	Impurity Function of Gini Index Measure	25
3.3.2	Sample Calculation of Gini Index	26
3.4	Applications of CART Algorithm	28
CHAPTER 4	SYSTEM DESIGN AND IMPLEMENTATION	30
4.1	Overview of The System	30
4.2	Flow of the Proposed System	30
4.2.1	Dataset Description	32
4.3	Implementation of the Proposed System	35
4.4	In Depth Analysis of Generated Rules	38
4.5	Accuracy Measure	41
4.5.1	Confusion Matrix for Imbalanced Classification	41
4.6	Analysis of System Performance	44
CHAPTER 5	CONCLUSION	45
5.1	Limitation of thesis	45
5.2	Further Extensions	46
REFERENCES		47
ANNEX		50

LIST OF FIGURES

Figures	Name of Figures	Pages
Figure 2.1	Process of Classification	7
Figure 2.2	Sample Process of Classification	8
Figure 2.3	Concept of Decision Tree	10
Figure 3.1	General Flow of a CART Algorithm	24
Figure 4.1	System Flow of Identifying Malaria Infected Result with CART Algorithm	31
Figure 4.2	System Flow of Identifying Malaria Treatment with Rule-Based Classifier	32
Figure 4.3	System Flow Diagram of the System from User Side	35
Figure 4.4	Log in Form	36
Figure 4.5	Malaria Diagnosis Data Entry Form	36
Figure 4.6	Modal box of Patient's Record	37
Figure 4.7	Modal box of Patient's Record who has G6PD Deficiency	37
Figure 4.8	Patient Dataset	38
Figure 4.9	Rules Generated from Decision Tree	39
Figure 4.10	Treatment Rules which Generated from Decision Tree	40
Figure 4.11	Estimating Accuracy with Holdout Method	41
Figure 4.12	Confusion Matrix for Positive and Negative Tuples	42

LIST OF TABLES

Tables	Name of Tables	Pages
Table 4.1	Major Attributes of Patients' Record	34
Table 4.2	Result of Classifier Accuracy with Different Amount of Sample Data	44

LIST OF EQUATIONS

Equations	Name of Equations	Pages
Equation 2.1	Calculation of the midpoint value	11
Equation 2.2	Calculation of the Euclidean distance	17
Equation 3.1	Calculation of the measurement of impurity	25
Equation 3.2	Calculation of the Gini Index of Binary Split	25
Equation 3.3	Calculation of the Reduction in Impurity	26
Equation 4.1	Calculation of the Sensitivity	42
Equation 4.2	Calculation of the Specificity	42
Equation 4.2	Calculation of the Accuracy	43

CHAPTER 1

INTRODUCTION

Malaria is a life-threatening disease. It is caused by parasites transmitted from human to human through the bites of infected female Anopheles mosquito of the genus Plasmodium. At the same time, malaria can be prevented and treated. Malaria can be contracted throughout the year. May to September is the peak season for transmission of malaria in the rainy season. Anybody can suffer malaria. At the same time, mobile and migrant population, people working at construction and plantation sites, forest goers.[13] In order to strengthen the malaria surveillance system, more modern computerized systems are needed.

1.1 Introduction

Nowadays, the systems for management of data, information and knowledge are offering new potential for improvement of different sectors. Data mining algorithms are mainly distinguished as descriptive or predictive. Classification is the data analysis task, where a model or classifier is constructed to classify the category class labels. Classification and regression tree (CART) algorithm is a model of binary partitioning decision tree which can evaluate relationships between different type of data values of malaria diagnosis. CART algorithm is useful to find the independent variable that creates the best similar group when splitting the data element.

Expert system is established to help people for making decision effectively. In medical sectors, expert system might be developed when the health staffs cannot be available as well as in the place of hard reach areas. CART algorithm is chosen to demonstrate the malaria diagnosis cases. The system uses patient status as class label, and this class label has two classes: positive and negative. To classify for malaria infection, sixteen significant symptoms are chosen for every patient. the patient records will be stored in database. If the patient status results with positive, monitoring treatment case will be carried out. Rule-based classifier is used to determine the treatment which is based on the age group of each malaria patient. The treatment system follows the National Malaria

Treatment Guideline which is co- published by Ministry of Health and World Health Organization.

1.2 Objectives Of the Thesis

This thesis is submitted to reflect the following objectives.

- (i) To classify the malaria infected patient without receiving laboratory test results.
- (ii) To support the health staffs for providing valuable information using optimal decision tree classifier.
- (iii) To identify early diagnosis of malaria that can benefit to give suitable treatment for decreasing mortality rate of malaria.
- (iv) To reduce the missing treatment by monitoring with Malaria National Treatment Guideline.
- (v) To understand how to split attribute and how to develop decision tree using CART algorithm.

1.3 Method of Study

Data mining is known as an inter-disciplinary subfield of computer science and basically is a computing process of discovering patterns in large data sets. It is considered as an essential process where intelligent methods are applied in order to extract data patterns. The decisions are given by the questions that. Obviously, the tree will not always perfectly fit the data such as this one. This system uses one of the classification methods of decision tree induction. Malaria Diagnosis and Treatment System can support the health staffs in making decision to diagnosis of malaria by using CART algorithm and identify the treatment according to the age group of each patient. CART algorithm used Gini index impurity function and can produce binary tree.

1.4 Organization of Thesis

The thesis was organized with five chapters.

Chapter 1 discusses the introduction of thesis. It includes objective of the thesis, method of study and organization of the thesis. **Chapter 2** describes data mining process, components of data mining, classification with their methods, and some application in medical field. **Chapter 3** includes decision tree learning, CART algorithm, types of CART variable, Gini index measure, example of Gini index calculation, and application of CART algorithm. **Chapter 4** contains overview of the system, design of the system, implementation of the system, and the result of accuracy. Finally, the conclusions, the advantages over the proposed system, the limitation and the future work of the proposed system are expressed in the **Chapter 5**.

CHAPTER 2

CLASSIFICATION IN DATA MINING

Data Mining is used to turn raw data into useful information. By using software to find patterns in large amounts of data, businesses and organization can develop more effective strategies. Data mining involves exploring large amounts of information and studying meaningful patterns and trends. It can analyze the relationships and patterns of data, based on user requests.

2.1 Data Mining Process

Data mining is gaining popularity in different research fields due to its boundless applications and approaches to mine the data in an appropriate method. Due to the changes, the current world obtaining, it is one of the optimal approaches for approximating the nearby future consequences.[1] Along with advanced research in healthcare data are available, but the main difficulty is how to cultivate the existing information into a useful practice. To clarify this issue, the concept of data mining is the best suited. Data mining have a great potential to enable healthcare systems to use data more efficiently and effectively.

Data mining is considered as a whole process consisting of different steps where in each step different data mining techniques can be used. Then, several steps of the data mining process and the possible data mining techniques that might be used in each step are tried to classify as follows.

- (i) Data gathering
- (ii) Data preprocessing
 - Data cleaning (to remove noise and inconsistent data)
 - Data integration (where multiple data sources may be combined)
 - Data selection (where data relevant to the analysis task are retrieved from database)

- Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

(iii) Data Modeling

(iv) Data Optimization

2.2 Components of Data Mining System

The architecture of a typical data mining system may have the following major components:[11]

(i) Database, data warehouse, World Wide Web, or other information repository

This is one or a set of databases, data warehouse, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

(ii) Database or data warehouse server

The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

(iii) Knowledge base

This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes values into different levels of abstraction.

(iv) Data mining engine

This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

(v) Pattern evaluation module

This employs interestingness measures and interacts with the data mining modules to focus the search on interesting patterns. It may use interestingness thresholds to filter out discovered patterns.

(vi) User interface

This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.

2.3 Process of Classification

Classification is the process of classifying a record. One simple example of classification is to check whether it is raining or not. The answer can either be yes or no. So, there is a particular number of choices. Sometimes there can be more than two classes to classify. Another process of data analysis is prediction. It is used to find a numerical output. Same as in classification, the training dataset contains the inputs and corresponding numerical output values. The algorithm derives the model or a predictor according to the training dataset.[11] The model should find a numerical output when the new data is given. Unlike in classification, this method does not have a class label. The model predicts a continuous-valued function or ordered value. Regression is generally used for prediction. Predicting the value of a house depending on the facts such as the number of rooms, the total area are examples for prediction.

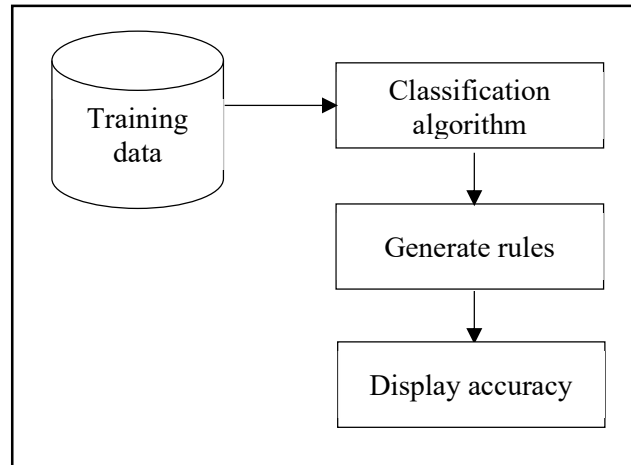
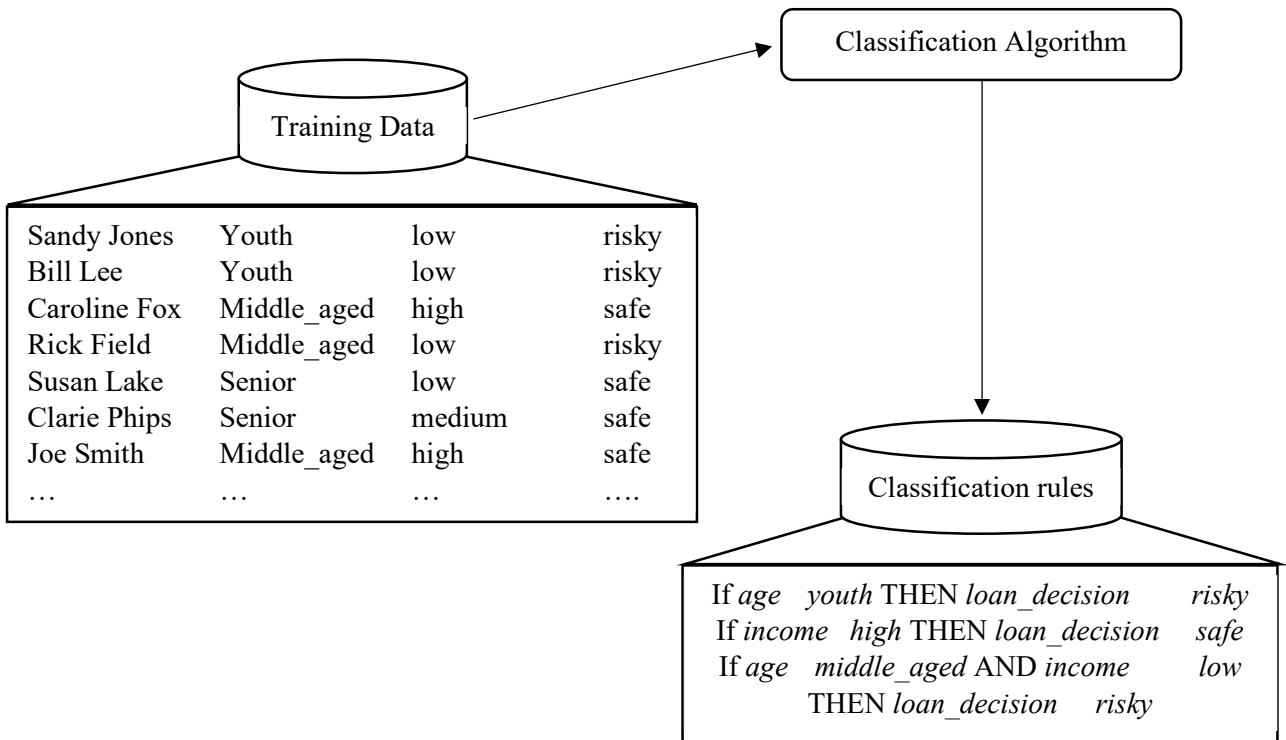


Figure 2.1 Process of Classification

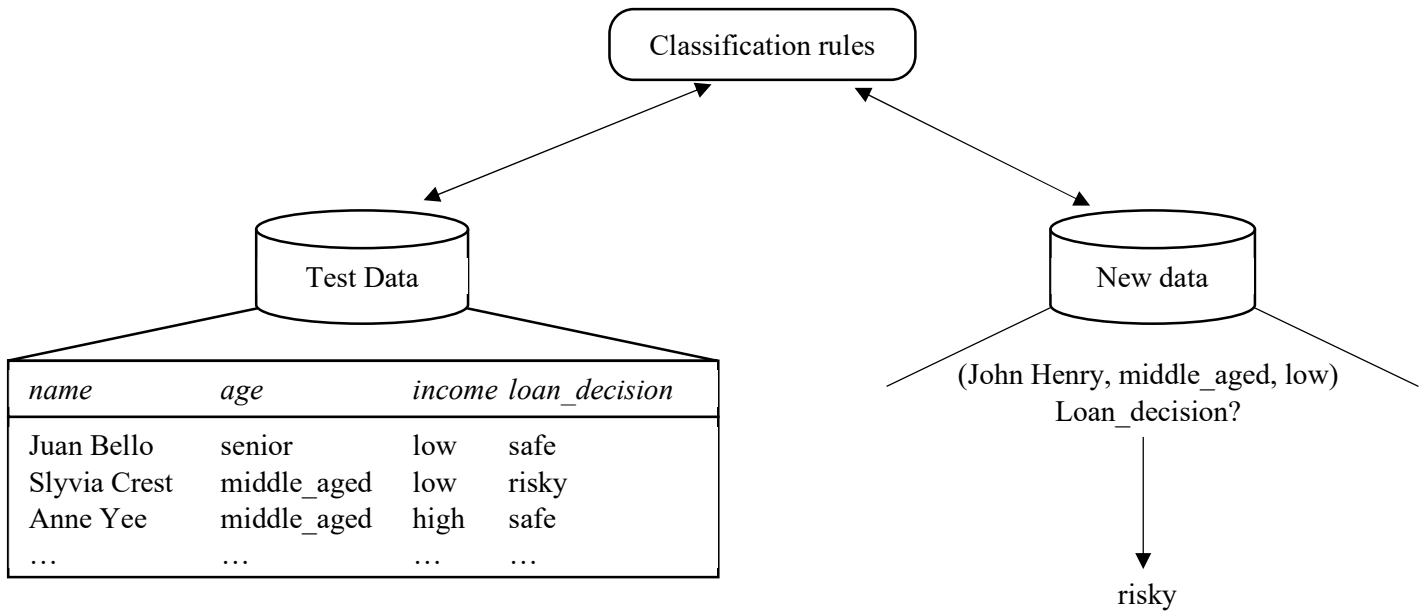
Data classification is a two-step process. In the first step, a classifier (or model) is built describing a predetermined set of data classes or concept. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or learning from a training set made up of database tuples and their associated class labels. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. On the context of classification, data tuples are also referred to as samples, examples, or objects. The data tuples analyzed to build the model collectively form the training dataset. The individual tuples making up the training set are referred to as training samples and are randomly selected from the population. Figure 2.1 describes the process of classification.

Since the class label of each training sample is provided, this step is also known as supervised learning (i.e., the learning of the classifier is supervised in that it is told to which class each training samples belongs). It contrasts with unsupervised learning (clustering), in which the class label of each training sample is not known, and the number or set of classes to be learned may not be known in advance.

Data classification consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data). The process is shown for the loan application data of Figure 2.2. The data are simplified for illustrative purposes.[11]



(a)



(b)

Figure 2.2 Sample Process of Classification (a) Learning (b) Classification

In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or “learning from” a training set made up of database tuples and their associated class labels. The class label attribute is discrete-valued and unordered. It is categorical (or nominal) in that each value serves as a category or class. The individual tuples making up the training set are referred to as training tuples and are randomly sampled from the database under analysis.[11] In the context of classification, data tuples can be referred to as samples, examples, instances, data points, or objects.

2.4 Method of Classification

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. For example, a classification model can be built to categorize bank loan applications as either safe or risky. Such analysis can help provide us with a better understanding of the data at large. Many classification methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large amounts of disk-resident data. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis.[11]

2.4.1 Classification with Decision Tree Induction

Decision tree algorithms (e.g., ID3, C4.5, and CART) were originally intended for classification. Decision tree induction constructs a flowchart like structure where each internal (non-leaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction. At each node, the algorithm chooses the best attribute to partition the data into individual classes. When decision tree induction is used for attribute subset selection, a tree is constructed from the given data. All attributes that do not appear in the tree are assumed to be irrelevant. The set of attributes appearing in the tree form the reduced subset of attributes.

During the construction of decision tree, the data is split into smaller subsets iteratively. At each iteration, choosing the most suitable independent variable is an important issue. The split which creates the most homogenous subsets with respect to the dependent variable should be chosen.[19] While choosing the independent variable, some attribute selection measures like information gain, Gini index etc. are used. Then, these splitting processes according to the measures continue until no more useful splits are found. In brief, decision tree technique is useful for classification problems and the most common types of decision tree algorithms are ID3, C4.5, and CART. Figure 2.3 refers to the concept of decision tree.

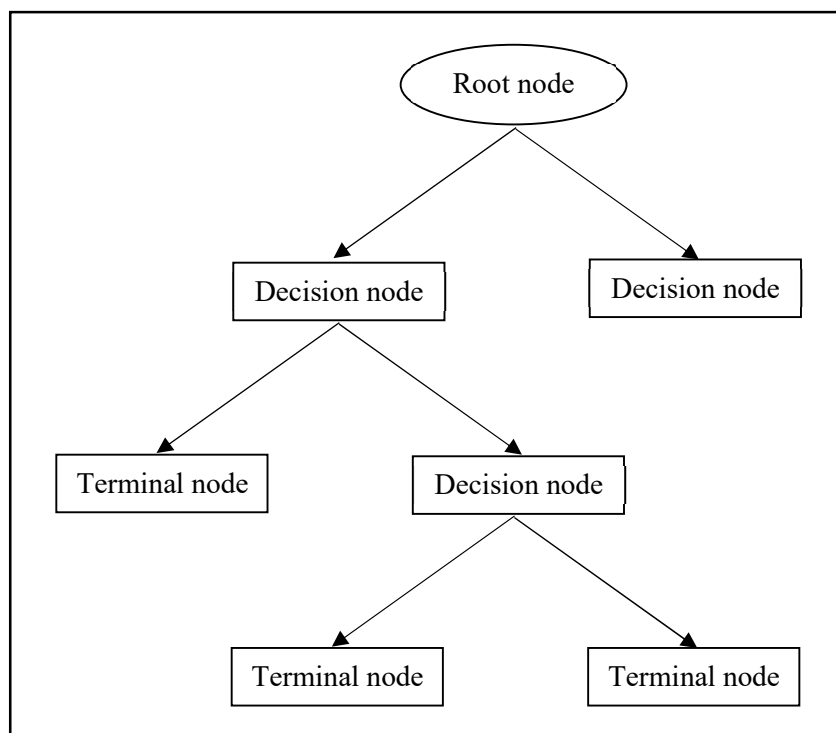


Figure 2.3 Concept of Decision Tree

An attribute selection measure is a heuristic for selecting the splitting criterion that best separates a given data partition, D , of class-labeled training tuples into individual classes. According to the outcome of the splitting criterion, each partition would be (i.e., all the tuples that fall into a given partition would belong to the same class). The best splitting criterion is the one that most closely results in such a scenario. Attribute selection measures are also known as splitting rules because they determine how the tuples at a given

node are to be split. The attribute selection measure provides a ranking for each attribute describing the given training tuples. The attribute having the best score for the measure (i.e., depending on the measure, either the highest or lowest score is chosen as the best) is chosen as the splitting attribute for the given tuples.

If splitting attribute is continuous-valued, either a split point or splitting subset must be determined as part of the splitting criterion. To determine the best split for attribute A , sort the values of A with the increasing order. Typically, the midpoint between each pair of adjacent values is considered as a possible split point. Therefore, given v values of A , then possible of $v-1$ splits are evaluated.[11]

For examples, the midpoint between the values of a_i and a_{i+1} of A is

$$\frac{a_i + a_{i+1}}{2} \quad 2.1$$

Where, a_i = the value of attribute, A in the position of i

a_{i+1} = the value of attribute, A in the position of $i+1$

If the values of A are sorted in advance, determining of the best split for A requires only one pass through the values. For each possible split point for A , the number of data partition D is two. Then D is the set of tuples in D satisfying $A \leq \text{split point}$, D_2 is the split of tuples in D satisfying $A \geq \text{split point}$.

There are three popular attribute selection measures: **Information Gain**, **Gain ratio**, and **Gini index**. Information gain: The attribute with the highest information gain is chosen as the splitting attribute. This attribute minimizes the information needed to classify the tuples in the resulting partitions.

(i) Information gain

ID3 (Iterative Dichotomiser 3) uses information gain as its attribute's selection measure. To find out the information gain for each attribute and which argument comes first, then determine the entropy off the attributes. The entropy is a numeric value for estimating the amount of information gain if at one level is used a certain

attribute. ID3 builds the tree from the top down, with no backtracking. The attribute with the highest information gain is chosen as the splitting attribute.

(ii) Gain ratio

The information gain measure is biased towards tests with many outcomes. That is, it prefers to select attributes having a larger number of values. C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a split information value. The attributes with the maximum gain ratio are selected as the splitting attribute.

(iii) Gini Index

The Gini index is used in CART algorithm. It considers a binary split for each attribute. The attribute that maximizes the reduction in impurity or the minimum Gini index is selected as the splitting attribute. Gini index is biased towards multivalued attributes and has difficulty when the number of classes is large, but it has the equal partitions.[11]

2.4.2 Classification with Bayesian

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem, described next. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naïve." Various empirical studies of this classifier in comparison to

decision tree and neural network classifiers have found it to be comparable in some domains.

In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, in practice this is not always the case, owing to inaccuracies in the assumptions made for its use, such as class-conditional independence, and the lack of available probability data. Bayesian classifiers are also useful in that they provide a theoretical justification for other classifiers that do not explicitly use Bayes' theorem. For example, under certain assumptions, it can be shown that many neural network and curve-fitting algorithms output the maximum posteriori hypothesis, as does the naïve Bayesian classifier.[11]

2.4.3 Classification with Rule-based

Rules are a good way of representing information or bits of knowledge. A rule-based classifier uses a set of *IF-THEN* rules for classification. An *IF-THEN* rule is an expression of the form

IF condition *THEN* conclusion.

An example is rule R1,

R1: *IF* *age = youth* AND *student = yes* *THEN* *buys computer = yes*.

The “IF” part (or left side) of a rule is known as the rule antecedent or precondition. The “THEN” part (or right side) is the rule consequent.[5] In the rule antecedent, the condition consists of one or more attribute tests (e.g., *age = youth* and *student = yes*) that are logically ANDed. The rule's consequent contains a class prediction (in this case, predicting whether a customer will buy a computer). R1 can also be written as

R1: (*age = youth*) ^ (*student = yes*) → (*buys computer = yes*).

If the condition (i.e., all the attribute tests) in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied (or simply, that the rule is satisfied) and that the rule covers the tuple.

2.4.4 Classification with Backpropagation

Backpropagation is a neural network learning algorithm. The neural networks field was originally kindled by psychologists and neurobiologists who sought to develop and test computational analogs of neurons. Roughly speaking, a neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Neural network learning is also referred to as connectionist learning due to the connections between units.

Neural networks involve long training times and are therefore more suitable for applications where this is feasible. They require a number of parameters that are typically best determined empirically such as the network topology or “structure.” Neural networks have been criticized for their poor interpretability. For example, it is difficult for humans to interpret the symbolic meaning behind the learned weights and of “hidden units” in the network. These features initially made neural networks less desirable for data mining.

The backpropagation algorithm performs learning on a *multilayer feed-forward* neural network. It iteratively learns a set of weights for prediction of the class label of tuples. A multilayer feed-forward neural network consists of an *input layer*, one or more *hidden layers*, and an *output layer*. Each layer is made up of units. The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of “neuronlike” units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network’s prediction for given tuples.[11]

2.4.5 Classification with Support Vector Machine

Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines can be defined as systems which

use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory.

Support Vector Machine becomes famous when, using pixel maps as input; it gives accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task. It is also being used for many applications, such as handwriting analysis, face analysis and so forth, especially for pattern classification and regression-based applications. The foundations of Support Vector Machines (SVM) have been developed by Vapnik and gained popularity due to many promising features such as better empirical performance. The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior, to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks. SRM minimizes an upper bound on the expected risk, whereas ERM minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems.

Early machine learning algorithms aimed to learn representations of simple functions. Hence, the goal of learning was to output a hypothesis that performed the correct classification of the training data and early learning algorithms were designed to find such an accurate fit to the data. The ability of a hypothesis to correctly classify data not in the training set is known as its generalization. SVM performs better in term of not over generalization when the neural networks might end up over generalizing easily. Another thing to observe is to find where to make the best trade-off in trading complexity with the number of epochs; the illustration brings to light more information about this.[6]

2.4.6 Classification with Association

Association rule learning is a machine learning method for discovering interesting relationships between variables in large databases. It is designed to detect strong rules in the database based on some interesting metrics. For any given multi-item transaction, association rules aim to obtain rules that determine how or why certain items are linked.

Association rules are created by searching for information on common if-then patterns and using specific criteria with support and trust to define what the key relationships are. They help to show the frequency of an item in a given data since confidence is defined by the number of times an if-then statement is found to be true. However, a third criterion called lift is often used to compare expected and actual confidence. Lift shows how many times the if-then statement was predicted to be true. Create association rules to compute itemset based on data created by two or more items. Association rules usually consist of rules that are well represented by the data.

There are different types of data mining techniques that can be used to find out the specific analysis and result like Classification analysis, Clustering analysis, and multivariate analysis. Association rules are mainly used to analyze and predict customer behavior. An associative classifier is a supervised learning model that uses association rules to assign a target value. The model generated by the association classifier and used to label new records consists of association rules that produce class labels. Therefore, they can also be thought of as a list of “if-then” clauses: if a record meets certain criteria (specified on the left side of the rule, also known as antecedents), it is marked (or scored) according to the rule’s category on the right. Most associative classifiers read the list of rules sequentially and apply the first matching rule to mark new records. Association classifier rules inherit some metrics from association rules, such as Support or Confidence, which can be used to rank or filter the rules in the model and evaluate their quality.[3]

2.4.7 Classification with K-Nearest-Neighbor

Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n -dimensional space. In this way, all the training tuples are stored in an n -dimensional pattern space. When given an unknown tuple, a k -nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k “nearest neighbors” of the unknown tuple.

“Closeness” is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, $X_1 = (x_{11}, x_{12}, \dots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \dots, x_{2n})$, is

$$\text{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2} \quad 2.2$$

For k-nearest-neighbor classification, the unknown tuple is assigned the most common class among its k-nearest neighbors. When $k \geq 1$, the unknown tuple is assigned the class of the training tuple that is closest to it in pattern space. Nearest-neighbor classifiers can also be used for numeric prediction, that is, to return a real-valued prediction for a given unknown tuple. In this case, the classifier returns the average value of the real-valued labels associated with the k-nearest neighbors of the unknown tuple.[11]

2.4.8 Classification with Case-Based Reasoning

Case-based reasoning (CBR) classifiers use a database of problem solutions to solve new problems. Unlike nearest-neighbor classifiers, which store training tuples as points in Euclidean space, CBR stores the tuples or “cases” for problem solving as complex symbolic descriptions. Business applications of CBR include problem resolution for customer service help desks, where cases describe product-related diagnostic problems. CBR has also been applied to areas such as engineering and law, where cases are either technical designs or legal rulings, respectively. Medical education is another area for CBR, where patient case histories and treatments are used to help diagnose and treat new patients.

When given a new case to classify, a case-based reasoner will first check if an identical training case exists. If one is found, then the accompanying solution to that case is returned. If no identical case is found, then the case-based reasoner will search for training cases having components that are similar to those of the new case. Conceptually, these training cases may be considered as neighbors of the new case. If cases are represented as graphs, this involves searching for subgraphs that are similar to subgraphs within the new case. The case-based reasoner tries to combine the solutions of the neighboring training cases to propose a solution for the new case. If incompatibilities arise

with the individual solutions, then backtracking to search for other solutions may be necessary. The case-based reasoner may employ background knowledge and problem-solving strategies to propose a feasible combined solution.[11]

2.5 Other Classification Method

There are multiple types of classification algorithms, each with its unique functionality and application. All of those algorithms are used to extract data from a dataset. Which application use for a particular task depends on the goal of the task and the kind of data you need to extract.

2.5.1 Classification with Genetic Algorithms

Genetic algorithms attempt to incorporate ideas of natural evolution. In general, genetic learning starts as follows. An initial population is created consisting of randomly generated rules. Each rule can be represented by a string of bits.

For example, suppose that samples in a given training set are described by two Boolean attributes, A_1 and A_2 , and that there are two classes, C_1 and C_2 . The rule “*IF A_1 AND NOT A_2 THEN C_2* ” can be encoded as the bit string “100,” where the two leftmost bits represent attributes A_1 and A_2 , respectively, and the rightmost bit represents the class. Similarly, the rule “*IF NOT A_1 AND NOT A_2 THEN C_1* ” can be encoded as “001.” If an attribute has k values, where $k > 2$, then k bits may be used to encode the attribute’s values. Classes can be encoded in a similar fashion.

Based on the notion of survival of the fittest, a new population is formed to consist of the *fittest* rules in the current population, as well as *offspring* of these rules. Typically, the *fitness* of a rule is assessed by its classification accuracy on a set of training samples. Offspring are created by applying genetic operators such as crossover and mutation. In *crossover*, substrings from pairs of rules are swapped to form new pairs of rules. In *mutation*, randomly selected bits in a rule’s string are inverted.[11]

The process of generating new populations based on prior populations of rules continues until a population, P , evolves where each rule in P satisfies a prespecified fitness threshold. Genetic algorithms are easily parallelizable and have been used for classification

as well as other optimization problems. In data mining, they may be used to evaluate the fitness of other algorithms.

2.5.2 Classification with Rough Set Approach

Rough set theory can be used for classification to discover structural relationships within imprecise or noisy data. It applies to discrete-valued attributes. Continuous-valued attributes must therefore be discretized before its use. Rough set theory is based on the establishment of equivalence classes within the given training data. All the data tuples forming an equivalence class are indiscernible, that is, the samples are identical with respect to the attributes describing the data. Given real-world data, it is common that some classes cannot be distinguished in terms of the available attributes. Rough sets can be used to define such classes approximately or roughly.

Rough sets can also be used for attribute subset selection or feature reduction, where attributes that do not contribute to the classification of the given training data can be identified and removed and relevance analysis where the contribution or significance of each attribute is assessed with respect to the classification task.[11]

2.5.3 Classification with Fuzzy Set Approach

Rule-based systems for classification have the disadvantage that they involve sharp cutoffs for continuous attributes. Fuzzy set theory is also known as possibility theory. It was an alternative to traditional two-value logic and probability theory. It lets us work at a high abstraction level and offers a means for dealing with imprecise data measurement. Fuzzy logic uses truth values between 0.0 and 1.0 to represent the degree of membership that a certain value has in a given category. Each category then represents a fuzzy set. Fuzzy set theory is useful for data mining systems performing rule-based classification. It provides operations for combining fuzzy measurements.

Given a tuple to classify, more than one fuzzy rule may apply. Each applicable rule contributes a vote for membership in the categories. Typically, the truth values for each predicted category are summed, and these sums are combined. Several procedures exist for translating the resulting fuzzy output into a defuzzified or crisp value that is returned by

the system. Fuzzy logic systems have been used in numerous areas for classification, including market research, finance, health care, and environmental engineering.[11]

2.6 Review on Data Mining in Healthcare

The healthcare sector requires data mining in discovery of knowledge and finding patterns for decision making. Data mining is the most advancing field of study which requires finding useful and meaningful details from a large data. Health data requires analytical methodology in identifying vital information that are used for decision making. Detection, prevention and management of diseases including fraud in the health insurance, reduce spending in the solution of medical care are some of the importance of data mining. It also helps researchers to make effective healthcare policies, develop recommendation systems and health profiles for patients. Volumes of data are generated in the healthcare industry that needs a database system to be stored for proper diagnosis and treatment of patients. The volumes of data are complicated and complex to analyze so as to make meaningful decision about the patient health status. Data can consist of the cost for treatment, hospital, medical claims, patients, doctor, history etc. due to the complexity of the data, data mining tools is necessary for analyzing and discovering knowledge from the data to enhance the processes of the patient and management. The result of using data mining in healthcare sector is for classifying diseases of the patients and to assist in treatment and management of diseases. It helps to predict the duration of admission of patients in hospital, diagnosis of patients and accurate management information system. Currently, technologies and data mining techniques help to reduce spending and evaluate the features that are responsible for diseases.[11]

Application of data mining have relevant use in healthcare. It is important to collect, store, prepared and mined data, to make healthcare data clean and correct. Clinical practices and standardization of distributing data across organizations to help in healthcare data mining technologies. Data mining promises great benefits in healthcare sector. The slow advance of technology and complexity of the volume of data make implementation of data mining strategies difficult. Till date, data mining in most part remains an academic practice. Data mining techniques had been used by academicians such as Neural Networks,

decision trees, Support Vector machine, Naïve Bayes and genetic algorithm to write and publish research papers.

Data mining processes can be fully or partially automated to analyze the volume of data that are uncertain such as cluster of data, anomaly detection or outliers and data dependencies. Input data of patients are collected into the database based on the dataset features which are further used for analysis in diagnosis in order to obtain more accurate prediction outcome for decision making. That data mining process are data collection preparation, data collection, data preprocessing, and data transformation but do not include knowledge extraction and evaluation steps.

Healthcare data are stored in electronic format all over the world in health organization. The format of the data contains patient's details which are of vast data. Due to the increase in in data, there exist complexity and complications. It can be worrisome when using traditional methods in analysis this set of data to generate meaning knowledge from it. The field of mathematics, computer and statistics makes it easy to discover meaningful information from volumes of complex data which makes data mining to be of great benefits to the healthcare sector.[17]

Data mining extracts meaningful information from complexity of data which were in a raw form. Numerous benefits are provided with the use of data mining in healthcare such as detection of fraud, detection of abuse of drugs, proper diagnosing of patients, treatments, early detection of diseases, survivability of patients etc.

Data mining techniques have been applied by various researchers. Such techniques are classification, association, clustering etc. the techniques play a vital role in the healthcare industry to support decision making, proper diagnosis, selection of treatments and prediction.

CHAPTER 3

DECISION TREE USING CART ALGORITHM

3.1 Decision Tree Learning

Decision tree learning is a common method used in data mining. Decision trees are powerful and popular tools for classification and prediction. The attractiveness of decision trees is because, in contrast to neural networks, decision trees represent rules. Rules can readily be expressed so that humans can understand them. The goal of decision tree is to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of the input variables, there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A tree can be learned by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value to the predictions. Since the decision tree is built by given data, the data value and character will be more important. For example, the amount of data will affect the result of the tree building procedure. The type of attribute value will also affect the tree model.

Decision trees need two kinds of data: training and testing. Training data, which are usually the bigger part of data, are used for constructing trees. The more training data collected, the higher the accuracy of the results. The other group of data, testing, is used to get the accuracy rate and misclassification rate of the decision tree. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Decision trees are the basis of several commercial rule induction systems.[19]

3.2 Classification and Regression Tree (CART) Algorithm

Classification and Regression Trees (CART) is a machine learning method. It is a data mining procedure to present the results of a complex data set in the form of decision tree. Decision trees are then used to classify new data. CART methodology was developed by Breiman, Friedman, Olshen and Stone in their paper Classification and Regression Trees in 1984. For building decision trees, CART uses so-called learning sample a set of historical data with pre-assigned classes.

CART is a non-parametric technique that can select variables and their interactions that are the most important in determining an outcome or dependent variable. If an outcome variable is continuous, CART produces regression trees, if the variable is categorical, CART produces classification trees. CART can handle both type of explanatory variables: categorical and continuous ones. CART can produce only binary tree and handle missing values automatically by using surrogate/substitute splits. CART uses Gini index impurity function for train data. Missing values in variables can be estimated by using surrogate variables so that partial data can be used whenever possible within the tree.

CART is a robust data mining and data analysis tool that automatically searches for important patterns and relationships and quickly uncovers hidden structure even in highly complex data.

The most important objective of decision trees is to seek accurate and small models. There are many criteria to measure node impurity. For example, the CART uses Gini index and C4.5 uses information gain. In CART algorithm, Gini index is used to split on an attribute that can maximally reduce impurity of the node. Gini index also tends to favor tests that result in equal-sized partitions and purity in both partitions.[8]

CART algorithm is as follows:

Algorithm: CART Algorithm

Input: Data partition, D , which is a set of training tuples and their associated class labels.

Output: A decision tree.

Method:

1. At the parent node, search all the possible splits for each predictor.
2. Choose the best split using the smallest impurity criterion among all possible predictors.
3. Split at that attribute.
4. If each child node is the terminal node having same class, then assigned class label.
5. Else, let the child node be the parent node, go to 1.
6. If each child node is terminal node, then labeled with majority class in that data partition.
7. Classify each terminal node using classification rule.

Figure 3.1 General Flow of a CART Algorithm

In the step 1, search all possible binary splits to each attribute according to its value. In step 2, Gini index measures to find the best splitting attribute, i.e., the smallest Gini index value attribute among all attributes. After that, the given data split into binary partition on the attribute table in step 3. If the partition of data has the same class and assigned this class label to that child de or terminal node, in step 4. In step 5, if the partition of the data has different classes, then define that partition as parent node and do through step 1 to step 5 until the terminal node has reached. In step 6, if the child node is terminal

node and there is no further split, then labeled that node with the majority cs in that data partition. In step 7, use classification rule and classify each terminal node with class label.

3.3 Types of CART Variables

CART uses Gini index to consider a binary split for each attribute. For discrete-valued attribute such as categorical variables, having v distinct values $\{a_1, a_2, \dots, a_v\}$, there are 2^v possible subsets. For example, if anemia has two possible values, namely $\{\text{yes}, \text{no}\}$, the possible subsets are $\{\text{yes}, \text{no}\}$, $\{\text{yes}\}$, $\{\text{no}\}$, $\{\}$. Excluding the power set of $\{\text{yes}, \text{no}\}$ and the empty set from the consideration do not represent a split. Therefore, the best split on *anemia* is $2^v - 2$. It was based on the binary split on that attribute.

For continuous-valued attribute A , sort the value in increasing order. Typically, the midpoint between each pair of adjacent values is considered as a possible *split point*. The point with the minimum expected information requirement for A is selected as the split point for A . there are two possible outcomes from node corresponding to the conditions $A \leq \text{split_point}$ and $A \geq \text{split_point}$. [11]

3.3.1 Impurity Function of Gini Index Measure

For the next step of CART algorithm, Gini index is used for measuring attribute selection. The smallest impurity criterion or minimum Gini index value is chosen in this step and CART uses impurity function to select the best attribute. To measure the impurity of D , the data partition or the set of training tuples is as below.

$$\text{Gini}(D) = 1 - \sum_{i=1}^n (p_i)^2 \quad 3.1$$

In facts, p_i is the probability that a tuple in D belongs to class C_i . The sum is computed over n class. A weighted sum of the impurity of each resulting partition was computed. For each attribute, if a binary split on A partitions, D will be divided into D_1 and D_2 and the Gini index of D given that partitioning is

$$\text{Gini}_A(D) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2) \quad 3.2$$

The attribute A maximizes the reduction in impurity is –

$$\Delta\text{Gini}(A)=\text{Gini}(D)-\text{Gini}_A(D) \quad 3.3$$

The attribute that maximizes the reduction in impurity is selected as the splitting attribute.[11]

3.3.2 Sample Calculation of Gini Index

CART uses Gini index impurity method. The sample data is shown as Annex A and it is used to calculate Gini index. For example, when the sample data has 23 records and 14 attributes, searching all possible binary splits to each attribute according to its value.

Step 1: Search all possible binary splits to each attribute according to its values. For example, in attribute *sudden_fever*, there are two attributes, namely $\{yes,no\}$ and 2^2 possible subsets $\{yes,no\}$, $\{yes\}$, $\{no\}$, $\{ \}$. To determine possible binary splits on *sudden_fever*, there is 2^2-2 possible ways to form two partitions of the data, based on binary split on the attribute $\{yes\}$ and $\{no\}$.

Step 2: Apply Gini index measure to each attribute split and find the best splitting attribute or minimum Gini index value attribute as follows:

Use Equation 3.1 to this data and find the purity of given data.

$$\text{Gini}(D) = 1 - (14/23)^2 - (9/23)^2 = 0.4764$$

After applying Equation 3.2, to search the Gini index values of each attribute are as follows:

$$\begin{aligned} \text{Gini}_{\text{sudden_fever}\{yes,no\}}(D) &= 16/23 (1 - (11/16)^2 - (5/16)^2) + 7/23 (1 - (3/7)^2 - (4/7)^2) \\ &= 0.4479 \end{aligned}$$

$$\begin{aligned} \text{Gini}_{\text{sudden_fever}\{no,yes\}}(D) &= 7/23 (1 - (3/7)^2 - (4/7)^2) + 16/23 (1 - (11/16)^2 - (5/16)^2) \\ &= 0.2400 \end{aligned}$$

$$\text{Gini}_{\text{headache}}(D) = 14/23 (1 - (10/14)^2 - (4/14)^2) + 9/23 (1 - (4/9)^2 - (5/9)^2)$$

$$= 0.4417$$

$$\text{Gini}_{\text{bleeding (D)}} = 9/23 (1 - (5/9)^2 - (4/9)^2) + 14/23 (1 - (9/14)^2 - (5/14)^2)$$

$$= 0.4727$$

$$\text{Gini}_{\text{muscle_pain (D)}} = 14/23 (1 - (10/14)^2 - (4/14)^2) + 9/23 (1 - (4/9)^2 - (5/9)^2)$$

$$= 0.4417$$

$$\text{Gini}_{\text{vomitting (D)}} = 16/23 (1 - (9/16)^2 - (7/16)^2) + 7/23 (1 - (5/7)^2 - (2/7)^2)$$

$$= 0.4288$$

$$\text{Gini}_{\text{diarrhea (D)}} = 12/23 (1 - (9/12)^2 - (3/12)^2) + 11/23 (1 - (5/11)^2 - (6/11)^2)$$

$$= 0.4328$$

$$\text{Gini}_{\text{weakness (D)}} = 12/23 (1 - (8/12)^2 - (4/12)^2) + 11/23 (1 - (6/11)^2 - (5/11)^2)$$

$$= 0.4690$$

$$\text{Gini}_{\text{jaundice (D)}} = 10/23 (1 - (8/10)^2 - (2/10)^2) + 13/23 (1 - (6/13)^2 - (7/13)^2)$$

$$= 0.4201$$

$$\text{Gini}_{\text{shock (D)}} = 12/23 (1 - (10/12)^2 - (2/12)^2) + 11/23 (1 - (4/11)^2 - (7/11)^2)$$

$$= 0.3663$$

$$\text{Gini}_{\text{urination_loss (D)}} = 11/23 (1 - (8/11)^2 - (3/11)^2) + 12/23 (1 - (6/12)^2 - (6/12)^2)$$

$$= 0.4506$$

$$\text{Gini}_{\text{dyspnea (D)}} = 16/23 (1 - (11/16)^2 - (5/16)^2) + 7/23 (1 - (3/7)^2 - (4/7)^2)$$

$$= 0.4479$$

$$\text{Gini}_{\text{convulsion (D)}} = 14/23 (1 - (13/14)^2 - (1/14)^2) + 9/23 (1 - (1/9)^2 - (8/9)^2)$$

$$= 0.1580$$

$$\text{Gini}_{\text{anemia}}(D) = 13/23 (1 - (11/13)^2 - (2/13)^2) + 10/23 (1 - (3/10)^2 - (7/10)^2)$$

$$= 0.3298$$

$$\text{Gini}_{\text{hemoglobinuria}}(D) = 15/23 (1 - (12/15)^2 - (3/15)^2) + 8/23 (1 - (2/8)^2 - (6/8)^2)$$

$$= 0.3391$$

Applying equation 3.3 to each attribute value and attribute *convulsion* has least impurity value and becomes root node.

$$\Delta\text{Gini}(\text{convulsion}) = \text{Gini}(D) - \text{Gini}_{\text{convulsion}}(D) = 0.4764 - 0.1580 = 0.3184$$

3.4 Applications of CART Algorithm

Classification and regression tree algorithm has been used in many applications areas, such as medicine, manufacturing and production, financial analysis, astronomy and molecular biology. The following are the review articles of CART algorithm.

Elia Georgiana Petre discussed CART algorithm with weather prediction domain. They used meteorological data of China between 2002 and 2005 and tried to forecast the future temperature values in Hong Kong. This statistical information about the data is used to generate the decision mode.

T. Santhanam and Shyam Sundaram used CART algorithm, one of the data mining modeling techniques, to examine the blood donor classification. The blood transfusion dataset (UCI ML repository) is based on donor database of Blood Transfusion Service Center in Hsin Chu City in Taiwan. In their work, CART algorithm creates a classification model with a combination of input parameters and considers the regular blood donor classification in India. The goal is to classify either Regular Voluntary Donor (RVD), a voluntary nonremunerated blood donor, who donates blood on a regular basis without any break for a longer duration between two donations or not RVD.

Helio Radke Bittencourt and Robin Thomas Clarke proposed CART algorithm in classification of remotely sensed digital images. They present a brief introduction to binary decision trees and show results obtained in the classifying Landsat-TM in southern Brazil and AVIRUS digital images. In Land-TM digital images, there are three classes of surface characteristic: namely water, cultivated land and natural vegetation. The AVIRUS image of a rural area in the state of Indiana, USA consists of three classes: corn notill, soybean notill and soybean minimum tillage. The training sample of pixels was used to construct the tree.

CHAPTER 4

SYSTEM DESIGN AND IMPLEMENTATION

Malaria is a potentially life-threatening disease. At the same time, malaria is preventable and curable. Early diagnosing can benefit to give suitable treatment for decreasing mortality rate of malaria. Users of this system might classify the malaria infected patient without receiving laboratory test results. In this system, users can keep patient records well-organized.

4.1 Overview of the System

In this system, the user needs to be collected training and testing data first. Dataset of malaria patients from Paletwa Township, Chin State of 2017 are used for diagnosis of malaria and 2,917 records of patients are stored in patient database. In the next step, the decision tree could be developed after calculating the impurity functions with CART algorithm. As a result, the rules will be generated that derived from the decision tree. In the step of testing, data are classified with rules to produce target class values are matched with identified values. The result of malaria positive or negative will be displayed to the user at the final stage.

According to the presented result, the treatment can be decided with the age group of each patient. This proposed system was also developed *IF-THEN* decision tree after identifying the malaria infection result of recorded patients. The rule-based classifier was used to generate the suitable treatment of malaria positive patients. The treatment is referencing from National Malaria Treatment Guideline which is co-published by Ministry of Health and World Health Organization and it's still active nowadays.

4.2 Flow of the Proposed System

This system is implemented to classify for diagnosis malaria patient. the system includes two processes. First, decision tree is developed, and rules are generated to identify malaria patients. Figure 4.1 shows the system design of classifying malaria patients with CART algorithm.

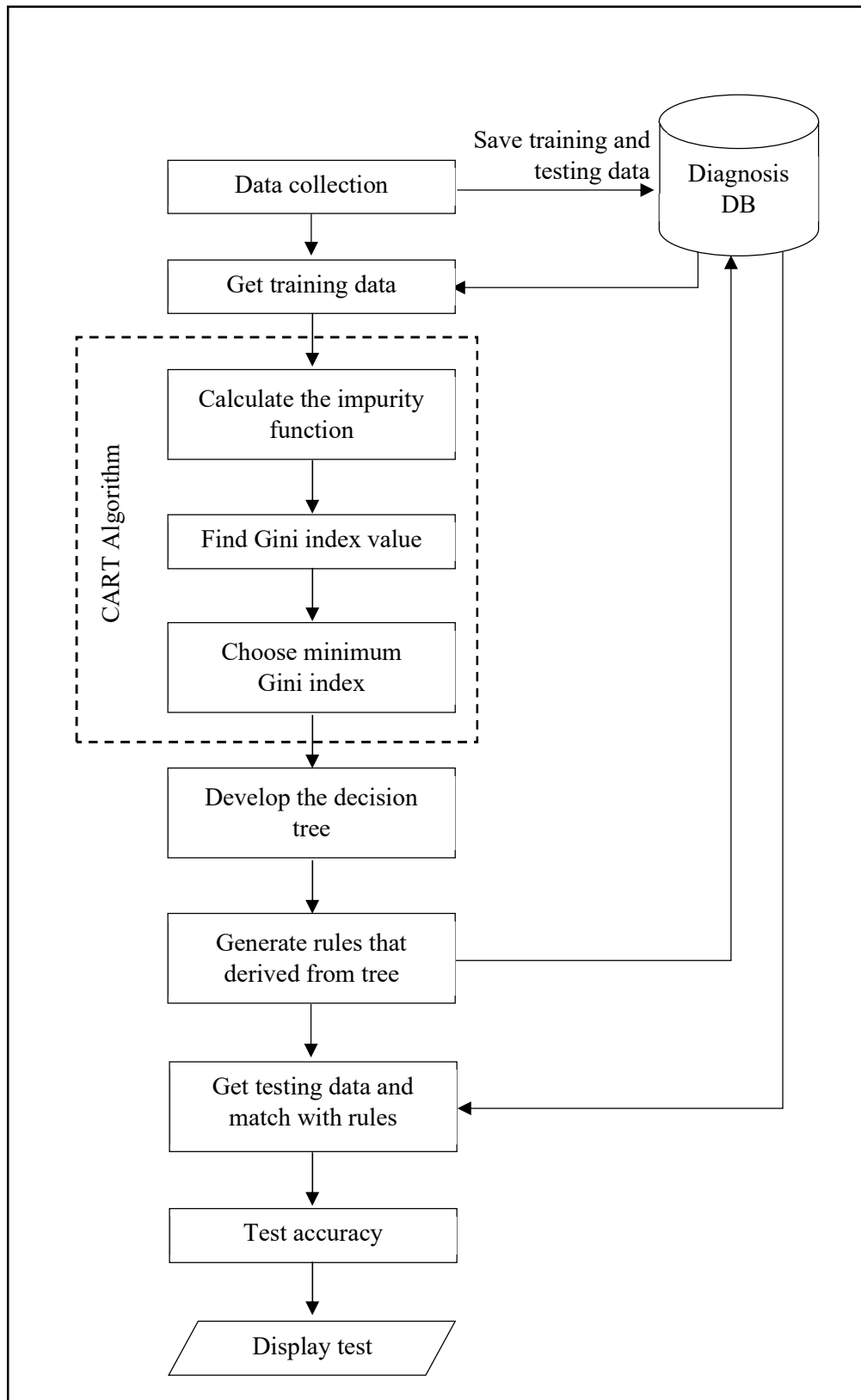


Figure 4.1 System Flow of Identifying Malaria Infected Result with CART Algorithm

Second, the correct treatment will be given to malaria positive patients according to National Malaria Treatment Guideline. Rule-based classifier is used to decide what treatment to be given. The flow of identifying malaria treatment is shown in Figure 4.2.

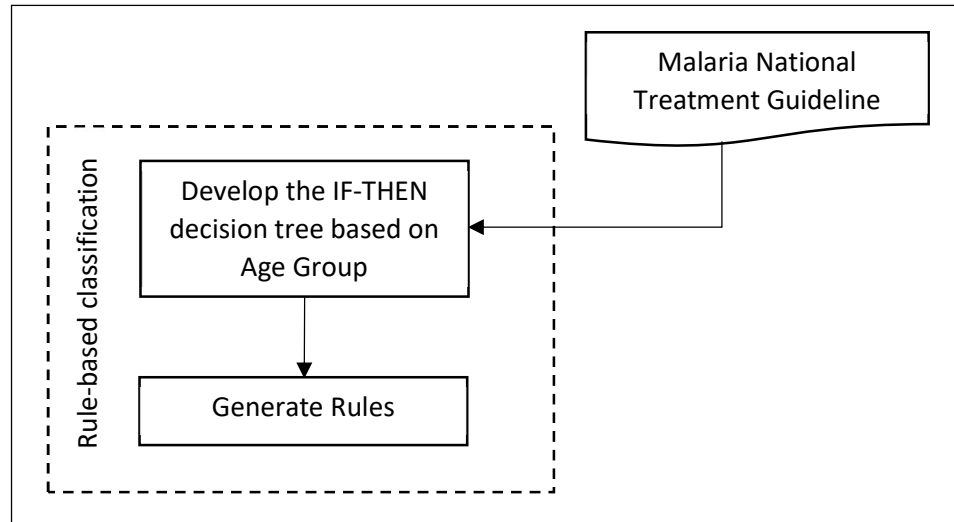


Figure 4.2 System Flow of Identifying Malaria Treatment with Rule-Based Classifier

4.2.1 Dataset Description

The epidemiology of malaria in Myanmar is highly complex. Symptoms of malaria include fever and flu-like illness, including shaking chills, headache, muscle aches, and tiredness. Nausea, vomiting, and diarrhea may also occur. Malaria may cause anemia and jaundice (yellow coloring of the skin and eyes) because of the loss of red blood cells. If not promptly treated, the infection can become severe and may cause kidney failure, seizures, mental confusion, coma, and death.

In this system, 16 attributes are used for predicting the diagnosis of malaria. The descriptions of each attribute are as follows:

- (i) Sudden fever (Immediately increase temperature about 39C)
- (ii) Vomiting (Persistent vomiting)
- (iii) Weakness (Cannot sit un-aid, cannot stand un-aid, cannot walk un-aid)

- (iv) Anemia (Body does not have enough red blood cells and is unable to deliver enough oxygen around the body)
- (v) Jaundice (Yellow coloration of the eyes)
- (vi) Headache
- (vii) Renal Failure (Failure to pass urine or passing a very small quantity)
- (viii) Hemoglobinuria (Black urine)
- (ix) Dyspnea (Difficulty in lying flat due to breathing problems)
- (x) Convulsions (a sudden violent, irregular movement of the body)
- (xi) Bleeding (Spontaneous bleeding)
- (xii) Shock (Circulatory collapse - shown by a feeble, very rapid pulse and cold limbs)
- (xiii) Pain (Muscles and joint pain)
- (xiv) Cough
- (xv) Diarrhea
- (xvi) Oversweating

The proposed system uses the significant malaria symptoms as the attributes and the malaria positive or negative as a class label. In fact, the patient who has been infected malaria, this data is recorded as a malaria history. Moreover, if a patient suffers G6PD deficiency, it will be recorded for taking care of accurate treatment and hospital referral.

The possible values of each attribute of patient are shown in Table 4.1.

Table 4.1 Major Attributes of Patients' Record

No	Symptoms	Datatype	Value
1	Patiend_ID	Integer	1,2,3,...
2	Sudden_fever	Boolean	Yes, No
3	Vomitting	Boolean	Yes, No
4	Weakness	Boolean	cannot sit un-aid, cannot stand un-aid, cannot walk un-aid
5	Anaemia	Boolean	Uncomplicated, Severe
6	Jaundice	Boolean	Yes, No
7	Headache	Boolean	Yes, No
8	RenalFailure	Boolean	Yes, No
9	Hemoglobinuria	Boolean	Yes, No
10	Dyspnea	Boolean	Yes, No
11	Convulsions	Boolean	Yes, No
12	Bleeding	Boolean	Yes, No
13	Shock	Boolean	Yes, No
14	Pain	Boolean	Yes, No
15	Cough	Boolean	Yes, No
16	Diarrhoea	Boolean	Yes, No
17	Oversweating	Boolean	Yes, No

Figure 4.3 refers to the system flow from the view of user side. After entering the record of a patient, the system classifies the symptoms of malaria diagnosis and generates the rules for that patient. Then the system identifies the patient is malaria positive or negative. If the result of patient is negative, the system will display the message of malaria negative with a modal box. If the result of patient is positive, the system will check for the appropriate treatment according to the Malaria National Treatment Guideline. The treatment given method is based on age group of patients and the treatment plan will be displayed at the end.

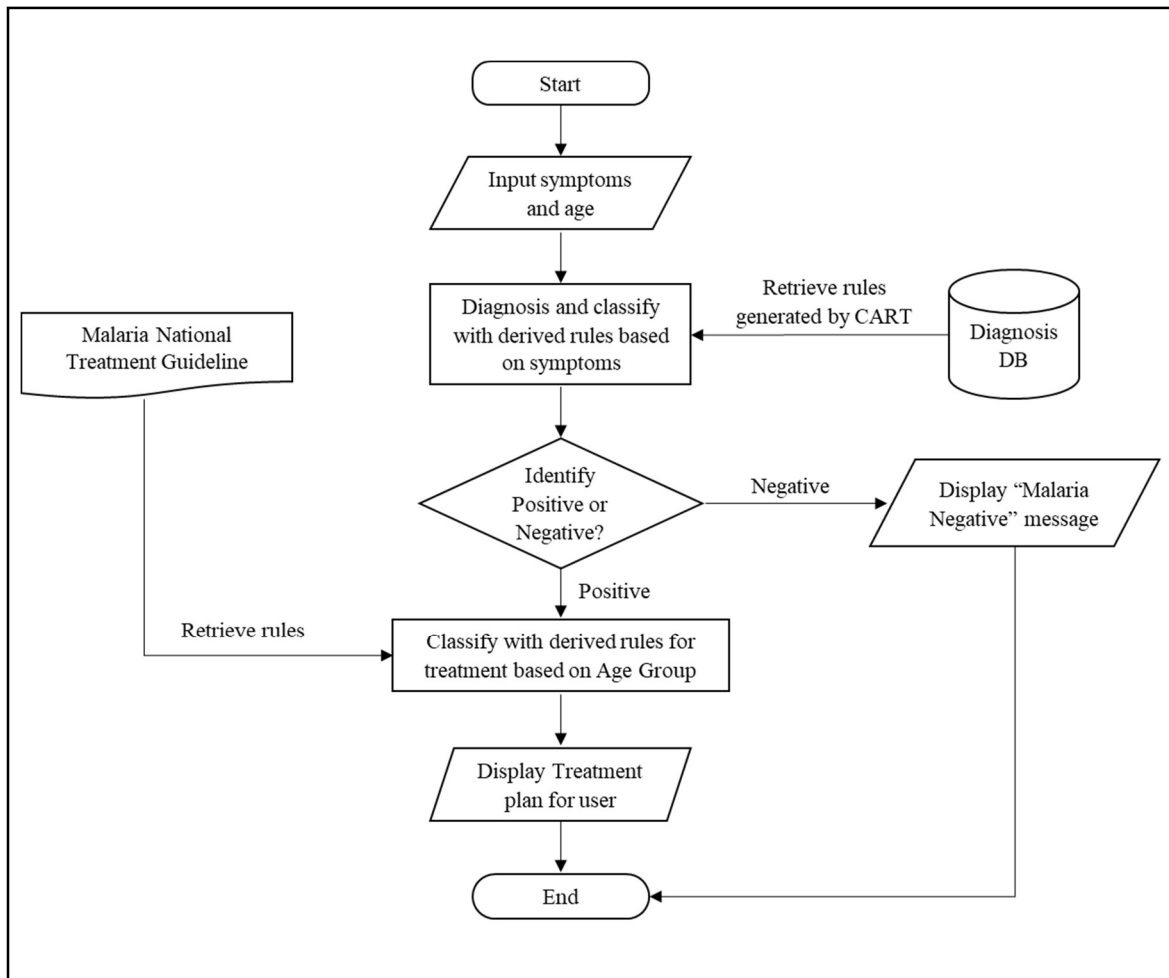


Figure 4.3 System Flow Diagram of the System from User Side

4.3 Implementation of the Proposed System

The system was developed with PHP and the user interface was designed with HTML and CSS. In the form of the system, there are two types of user roles such as administrator and user. When starting the application, the log in form will be displayed as Figure 4.4.

**Malaria Diagnosis
and Treatment System**

Email

Password

Remember Me

Login

Figure 4.4 Log in Form

Firstly, the user can make data entry to test malaria positive or not according to the significant symptoms. The system predicts malaria positivity results by means of six distinct symptoms out of sixteen signs and symptoms of malaria diseases. The data entry form is as shown in Figure 4.5.

Malaria Form **Logout**

Malaria Diagnosis Form

State /Divisions Township Address

Patient Name Date of Birth Gender Male Female

Pregnant Yes No History Yes No Suddenfever Yes No

Vomitting Yes No Anaemia Uncomplicated Severe Jaundice Yes No

Weakness cannot sit un-aid cannot stand un-aid cannot walk un-aid

Headache Yes No RenalFailure Yes No Hemoglobinuria Yes No

Dyspnea Yes No Convulsions Yes No Bleeding Yes No

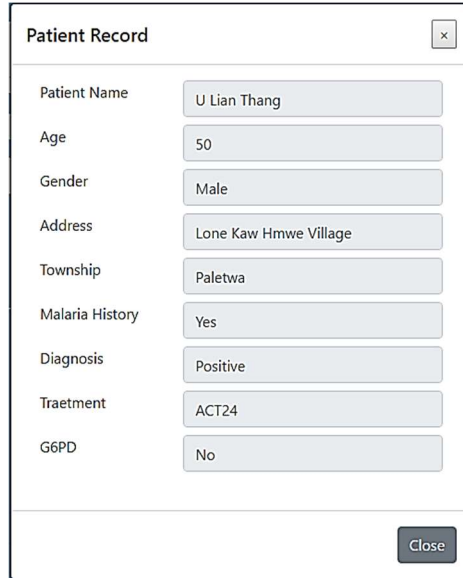
Shock Yes No Pain Yes No Cough Yes No

Diarrhoea Yes No OverSweating Yes No G6PD Yes No

save

Figure 4.5 Malaria Diagnosis Data Entry Form

After saving the completed entry record, the model box will be displayed to identify the patient was infected malaria diseases. Figure 4.6 is the sample modal box of generated patient's record.

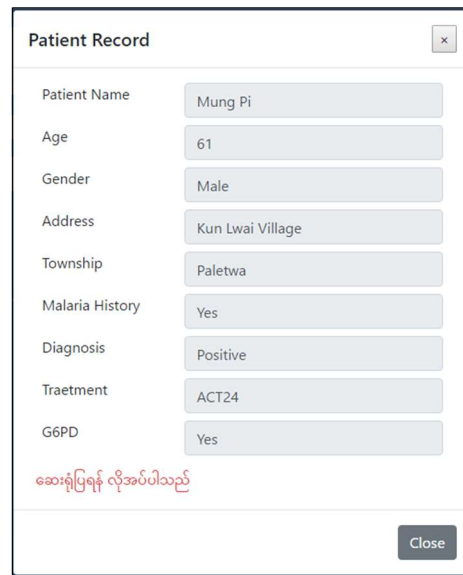


A modal box titled "Patient Record" with a close button (x) in the top right corner. It contains a list of patient information fields, each with a text input field:

Patient Name	U Lian Thang
Age	50
Gender	Male
Address	Lone Kaw Hmwe Village
Township	Paletwa
Malaria History	Yes
Diagnosis	Positive
Treatment	ACT24
G6PD	No

A "Close" button is located at the bottom right of the modal box.

Figure 4.6 Modal box of Patient's Record



A modal box titled "Patient Record" with a close button (x) in the top right corner. It contains a list of patient information fields, each with a text input field:

Patient Name	Mung Pi
Age	61
Gender	Male
Address	Kun Lwai Village
Township	Paletwa
Malaria History	Yes
Diagnosis	Positive
Treatment	ACT24
G6PD	Yes

Below the G6PD field, there is a red text label: "ဆေးရုံပြုရန် လိုအပ်ပါသည်" (Hospital treatment is required).

A "Close" button is located at the bottom right of the modal box.

Figure 4.7 Modal box of Patient's Record who has G6PD Deficiency

Glucose-6-phosphate dehydrogenase (G6PD) deficiency is relatively common in populations exposed to malaria. This deficiency appears to provide some protection from this infection, but it can also cause hemolysis after administration of some antimalarial drugs. The risk of drug-induced G6PD deficiency-related hemolysis depends on a number of factors including the G6PD variant, the drug and drug dosage schedule, patient status, and disease factors. If the patient is suffering G6PD deficiency, the system alerts to referral the patient for hospitalization. The referral message is displayed in Figure 4.7.

No	State	Township	Pt_Name	PMonth	PYear	Age_Year	G6PD	Pregnant	History	Suddenfever	Vomitting	Weakness
1	Chin	Tedim	aung aung	9	2022	5	1	1	1	1	0	1
2	Shan (East)	Mongyang	user II	9	2022	3	1	1	1	1	0	1
3	Chin	Thantlang	user three	9	2022	0	1	0	0	0	1	0
4	Shan (South)	Hopong	patient three	9	2022	8	0	0	0	1	2	0
5	Bago (West)	Pyay	Patient one	9	2022	7	1	1	1	1	1	1
6	Chin	Paletwa	htet ya ngwe aung	3	2017	0	1	7	0	0	1	0
7	Chin	Paletwa	nay soe hwin	3	2017	41	1	0	0	0	1	0
8	Chin	Paletwa	san ar awm	3	2017	3	1	7	0	0	1	0
9	Chin	Paletwa	may htan	3	2017	10	1	0	0	0	1	0
10	Chin	Paletwa	bu ai	3	2017	0	1	7	0	0	1	0

Figure 4.8 Patient Dataset

Every single entry will be recorded in dataset. Administrator can also search with desire keywords in the list of patients as well as export with an excel sheet as well as random patients list. Administrator can import the training data from Excel file to database and can view all patient records as shown in Figure 4.8.

4.4 In Depth Analysis of Generated Rules

The performance of system needs to be evaluated by testing with one-third of patient records. Before confirming the accuracy of measuring, the rules of diagnosis will be check whether it is working or not in real time usage.

```

<pre>
<?php

/*
 * Decision Tree Learning
 * Problem Solving: Predicting the output of instance (Malaria Function)
 * Additional: Testing Set for testing Accuracy
 * Source: http://pages.cs.wisc.edu/~shavlik/cs540/HWs/HW1.html
 */

include_once __DIR__ . '/vendor/autoload.php';

//tidy up data
$training_set = tidyMalariaData(__DIR__.'/malariadb.data');

$hepTree = new Jincongho\DecisionTree\DecisionTree;
$hepTree->setAttrNum(22)->addTrainingSet($training_set)->startTraining();

//print_r($hepTree->getTree());
//echo "\nGini Index per Attribute(column#, ordered by the ascending): \n";
//var_dump($hepTree->getGain());
//testing set
$testing_set = tidyMalariaData(__DIR__.'/malariadb_test.data');

$correct = 0;
$missing = 0;
foreach($testing_set as $set){
    if($hepTree->classify($set[0]) == $set[1]){
        $correct++;
    }else{
        $missing++;
    }
}

echo "\nRunning testing sets(", $correct + $missing, "):\n Correct: ",
$correct, "\nMissing: ",
$missing, "\nAccuracy: ", $correct/($correct+$missing);

function tidyMalariaData($file){
    $training_set = file($file);

    foreach ($training_set as $key => $value) {
        $filter = array_filter(explode(" ", $value), function($input){
            return strlen($input) > 0;
        });
        $filter = array_map('trim', $filter);
        $training_set[$key] = array(array_slice($filter, 2),
$filter[1]);
    }
    return $training_set;
}

```

Figure 4.9 Rules Generated from Decision Tree

Healthcare provider will examine and ask about the symptoms and malaria history. Depending on the type of parasite, symptoms can be mild. Some people do not feel sick for up to a year after the mosquito bite. The system provides sixteen symptoms to distinguish malaria positivity results. In facts, the eleven attributes such as sudden fever, vomiting, anemia, jaundice, renal failure, hemoglobinuria, dyspnea, convulsion, shock, muscle pain and over sweating were selected to pre-defined outputs. Among those symptoms, if the patient suffers six symptoms out of those, the diagnosis of his or her could be assumed as positive. Figure 4.9 represents to classify the derivation of rules.

The objective of the national antimalarial treatment policy is to provide safe and rapidly effective antimalarial treatment to all patients with malaria and to prevent the emergence and spread of drug resistance. The treatment given is ruled with age group of patients. In the output view, the end user can easily identify the age group by extracting rules from a decision tree. One rule is created for each path from the root to a leaf node. Figure 4.10 describes how to generate the output from decision tree.

```
$currentDate = date("d-m-Y");
$year = date('Y');
$month = date('m');

$age = date_diff(date_create($dob), date_create($currentDate));
$age_year = $age->format("%y");
switch ($age_year) {
    case $age_year >= 15:
        $age_group = 'ACT24';
        break;
    case $age_year < 14 && $age_year >= 10;
        $age_group = 'ACT18';
        break;
    case $age_year < 9 && $age_year >= 5;
        $age_group = 'ACT12';
        break;
    default:
        $age_group = 'ACT6';
        break;
}
```

Figure 4.10 Treatment Rules which Generated from Decision Tree

4.5 Accuracy Measure

Classification accuracy is the total number of correct predictions divided by the total number of predictions made for a dataset. As a performance measure, accuracy is inappropriate for imbalanced classification problems.

The main reason is that the overwhelming number of examples from the majority class (or classes) will overwhelm the number of examples in the minority class, meaning that even unskillful models can achieve accuracy scores of 90 percent, or 99 percent, depending on how severe the class imbalance happens to be.

In this system, hold out provides for estimating accuracy. The given data are randomly partitioned into two dependent sets, a training dataset and testing dataset. Typically, two-third of the data allocated to the training dataset and the remaining one-third is allocated to the test set. The training set is used to drive the model, whose accuracy is estimated with the test as in Figure 4.11.

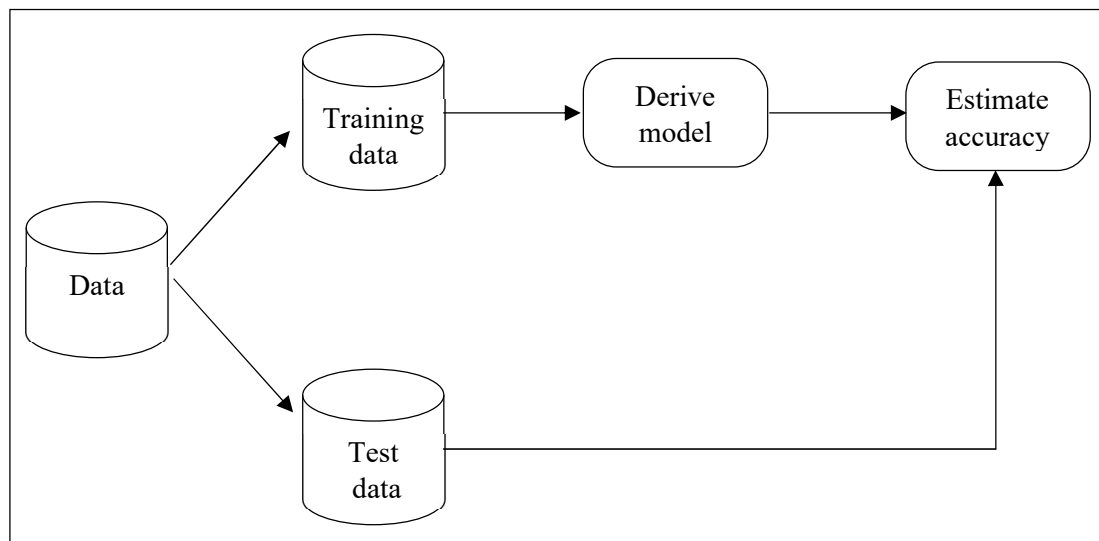


Figure 4.11 Estimating Accuracy with Holdout Method

4.5.1 Confusion Matrix for Imbalanced Classification

Before reflecting into precision and recall, it is important to review the confusion matrix. For imbalanced classification problems, the majority class is typically referred to

as the negative outcome such as negative result, and the minority class is typically referred to as the positive outcome such as positive test result.

The confusion matrix provides more insight into not only the performance of a predictive model, but also which classes are being predicted correctly, which incorrectly, and what type of errors are being made. The simplest confusion matrix is for a two-class classification problem, with negative (class 0) and positive (class 1) classes. In this type of confusion matrix, each cell in the table has a specific and well-understood name, summarized as follows:

		Predicted class	
		C ₁	C ₂
Actual class	C ₁	True positives	False negatives
	C ₂	False positives	True negatives

Figure 4.12 Confusion Matrix for Positive and Negative Tuples

The confusion matrix is a useful tool for analyzing how the classifier can recognize tuples of different classes. A confusion matrix for two classes (C₁ and C₂) is shown in figure 4.12. Given two classes, in terms of positive tuples, C₁ and negative tuples, C₂. The sensitivity and specificity measures can be used in accuracy measure.

$$\text{sensitivity} = \frac{\text{t-positive}}{\text{positive}} \quad (4.1)$$

$$\text{specificity} = \frac{\text{t-negative}}{\text{negative}} \quad (4.2)$$

where, sensitivity is true positive rate (the proportion of positive tuples that are correctly identified), specificity is true negative rate (the proportion of negative tuples that are correctly identified), t-pos is true Positive (the positive tuples that were correctly labeled by the classifier), -neg is true negative (the negative tuples that were correctly labeled by the classifier), pos is positive (the number of positive tuples) and neg is negative (the

number of negative tuples). It can be shown that accuracy is a function of sensitivity and specificity:

$$\text{accuracy} = \text{sensitivity} * \frac{\text{positive}}{(\text{positive} + \text{negative})} + \text{specificity} * \frac{\text{negative}}{(\text{positive} + \text{negative})} \quad (4.3)$$

When 917 records are corrected out of 1,000 test records and total positive case will be 798, then

$$\text{Sensitivity} = (917 - 798) / 798 = 0.14$$

$$\text{Specificity} = (917 - 119) / 119 = 6.70$$

$$\text{Accuracy} = 0.14 * (798 / (798 + 119)) + 6.70 * (119 / (798 + 119)) = 0.99$$

When 689 records are corrected out of 750 test records and total positive case will be 344, then

$$\text{Sensitivity} = (689 - 344) / 344 = 1.00$$

$$\text{Specificity} = (689 - 345) / 345 = 0.99$$

$$\text{Accuracy} = 1.00 * (344 / (344 + 345)) + 0.99 * (345 / (344 + 345)) = 0.99$$

When 459 records are corrected out of 500 test records and total positive case will be 258, then

$$\text{Sensitivity} = (459 - 258) / 258 = 0.77$$

$$\text{Specificity} = (459 - 201) / 201 = 1.28$$

$$\text{Accuracy} = 0.77 * (258 / (258 + 201)) + 1.28 * (201 / (258 + 201)) = 0.99$$

When 230 records are corrected out of 250 test records and total positive case will be 172, then

$$\text{Sensitivity} = (230 - 172) / 172 = 0.33$$

$$\text{Specificity} = (230 - 58) / 58 = 2.96$$

$$\text{Accuracy} = 0.33 * (172 / (172 + 58)) + 2.96 * (58 / (172 + 58)) = 0.99$$

When 91 records are corrected out of 100 test records and total positive case will be 64, then

$$\text{Sensitivity} = (91-64)/64 = 0.42$$

$$\text{Specificity} = (91-27)/27 = 2.37$$

$$\text{Accuracy} = 0.42 * (64 / (64 + 27)) + 2.37 * (27 / (64 + 27)) = 0.99$$

4.6 Analysis of System Performance

To determine Malaria Diagnosis System's performance the system was tested with different amount of sample data. The accuracy of system classifies increased with the number of sample data amount. The result of the test accuracy is described in Table 4.2.

Table 4.2 Result of Classifier Accuracy with Different Amount of Sample Data

No of sample records	No of test data records	No of corrected records	No of failed records	Classifier's accuracy (%)
1000	1000	917	83	99%
1000	750	689	61	99%
1000	500	459	41	99%
1000	250	230	20	99%
1000	100	91	9	99%

CHAPTER 5

CONCLUSION

This system is using the historical records of malaria patients and presenting the prediction of diagnosis by using data mining technique especially CART classification algorithm. This system is not intended to replace the medical experts but to help the experts to identify the diagnosis of malaria patient and to give correct treatments. This system can support some of the basic health staffs from hard-to-reach area to provide early diagnosis and giving treatment correctly.

CART is nonparametric so that this method does not require specification of any functional form. It can easily handle outliers and missing values using surrogate or substitute split. The records of malaria patient from Paletwa Township were used to demonstrate in this study. CART algorithm was demonstrated in malaria diagnosis and rule-based classifier was used to identify treatment of malaria positive patients. The extracted classification rules from decision tree are tested with various amounts of test data. As a result, the impact of large amount of sample data set on classifier's accuracy can be observed and this system can benefit in assisting basic health staffs with critical care decision making.

5.1 Limitation of thesis

In this thesis, 16 attributes which are distinct symptoms are used to classify the positivity of malaria. In conclusion, the system can support to perform early diagnosis and correct treatment. This system cannot be trusted completely to diagnosis of malaria. Computerized systems are well-adjusted to do repetitive tasks. These never get tired, bored or fatigued. Still, there can be failures of a computer system due to internal and external reasons. By the reference from Library of Medicine, malaria can hide in people with no symptoms, called "*asymptomatic malaria*". This system is not useful when some patients who suffer malaria have no significant symptom. This is one of the major limitations of this systems, but it will be useful most of the patients for diagnosing in common.

5.2 Further Extension

Techniques of data mining in diagnosis, tools used for data mining are reviewed. The approach of malaria diagnosis system can be extended by using more attributes and records. The integer valued attributes in this system can be split by using split point. The CART algorithm can also be used in regression tree when the target attribute variable is continuous. CART algorithm can also be used in other medical areas. The pruning of decision tree can be used to extend this system. The system concludes with health analytics stages and specific areas data mining application in healthcare. The knowledge will help to reduce unnecessary spending and make accurate decision from the volume and complexity of the health care data available.

REFERENCES

- [1] Akanksha A Kherdikar Kurlekar and Anusuya S. "A Study on Role of Data Mining in Research Methodology." *Indian Journal of Commerce & Management Studies*. March 3, 2011. www.sscholarshub.net.
- [2] Akash Ramaswamy, Chakrapani Mahabala, Sridevi Hanaganahalli Basavaiah, Animesh Jain, and Ravi Raj Singh Chouhan. "Asymptomatic malaria carriers and their characterization in hotpops of malaria at Mangalore." *National Center for Biotechnology Information*, June 2020.
- [3] *Associative Classification in Data Mining*. June 22, 2022. <https://www.geeksforgeeks.org/associative-classification-in-data-mining/>.
- [4] B, Nidhi. *Rule Based Data Mining Classifier: A Comprehensive Guide 101*. May 30, 2022. <https://hevodata.com/learn/rule-based-data-mining/>.
- [5] Biao Qin, Yin Xia, Sunil Prabhakar and Yicheng Tu. *A Rule-Based Classification Algorithm for Uncertain Data*. IEEE International Conference on Data Engineering, 2009.
- [6] Brownlee, Jason. "Machine Learning Mastery." *Machine Learning Mastery*. April 8, 2020. <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/>.
- [7] "Data Mining-Decision Tree Induction." n.d. https://www.tutorialpoint.com/data_mining/dm_dt.htm.
- [8] Deepankar. "Decision Tree with CART Algorithm." April 19, 2021. <https://medium.com>.
- [9] Hari, Vijaya. *Empirical Investigation of CART and Decision Tree Extraction form Neural Networks*. Ohio University, 2009.
- [10] Hnin Su Wai and Daw Myint Myint Maw. "Decision Support System For Mosquito Borne Disease." In *Decision Support System For Mosquito Borne Disease*, by Hnin

- Su Wai and Daw Myint Myint Maw. Yangon: University of Computer Studies, Yangon, 2009.
- [11] Jiawei Han, Micheline Kamber and Jian Pei. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers is an imprint of Elsevier, 2012.
- [12] Jonathan Abeles and David J Conway. "The Gini Coefficient As a Useful Measure of Malaria Inequality Among Populations." 2020. <https://doi.org/10.11.86/s12936-020-03489-x>.
- [13] Ministry of Health, Myanmar. *Guidelines for Diagnosis and Treatment of Malaria in Myanmar*. Malaria Manual Development Committee, 2002.
- [14] Mohd Mahmood Ali and Lakshmi Rajamani. "Decision Tree Inductiion: Data Classification using Height-Balanced Tree." *International Conference: Information and Knowledge Engineering*, n.d.
- [15] Njoku, Obinna Chilezie. *Decision tree and Their Application for Classification and Regression Problems*. Missouri State University, 2019.
- [16] Nyi Nyi Latt and Khin Moe Sann. "Diagnosis of Malarias by Using Reduct Generation Algorithm." In *Diagnosis of Malarias by Using Reduct Generation Algorithm*, by Nyi Nyi Latt and Khin Moe Sann. Hinthada: Computer University (Hinthada), 2011.
- [17] Ogundele I.O, Popoola O.L, Oyesola O.O and Orija K.T. "A Review on Data Mining in Healthcare." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 2018.
- [18] Parvez Ahmad, Saqib Qamar and Syed Qasim Afser Rizvi. "Techniques of Data Mining In Healthcare." *International Journal of Computer Applications*, 2015.
- [19] Seif, George. "A Guide to Decision Trees for Machine Learning and Data Science." November 30, 2018. <https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science>.

- [20] Soe Kalayar Naing and Nyein Nyein Myo. "Diagnosis of TB Disease by Using Decision Tree Induction." In *Diagnosis of TB Disease by Using Decision Tree Induction*, by Soe Kalayar Naing and Nyein Nyein Myo. Taungoo: Computer University (Taungoo), 2011.
- [21] Supajittree Boonnamnuay, Nittaya Kerdprasop and Kittisak Kerdprasop. "Classification and Regression Tree with Resampling for Classifying Imbalanced Data." *International Journal of Machine Learning and Computing*, 2018.
- [22] T.C.Olayinka and S.C.Chiemeke. "Predicting Paediatric Malaria Occurrence Using Classification Algorithm in Data Mining." *Journal of Advances in Mathematics and Computer Science*, 2019.
- [23] Win Min Thit, Jaramit Raewkungwal, Ngamphol Soonthornworasin, Nawanan Theera Amppumpunt, Boonchai Kijsanayotin, Saranath Lawpoolsri, Sid Naing and Wirichada Pan ngum. *Electronic Medical Record in Myanmar User Perceptions at Marie Stopes International Clinics in Myanmar*. Bangkok: Mahidol University, 2016.
- [24] XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh and Qiang Yang. *Top 10 Algorithms in Data Mining*. Survey, Knowl Inf Syst, 2007.

ANNEX

Annex 1: Sample Dataset

sudden_fever	Headache	Bleeding	muscle_pain	vomiting	diarrhea	Weakness	jaundice	shock	Urination_loss	dyspnea	Convulsion	anemia	Hemoglobinuria	Diagnosis
1	1	0	0	1	0	1	0	0	1	1	1	1	0	Malaria
1	0	1	1	1	0	0	0	1	1	0	1	1	1	Malaria
0	1	0	0	1	0	0	0	0	0	0	0	0	0	Chikungunya
1	1	1	1	1	1	0	0	0	0	1	0	0	1	Dengue
1	1	0	1	1	1	0	1	0	0	1	1	1	1	Malaria
1	0	0	0	1	0	1	1	0	0	0	0	0	0	Rift Valley fever
1	0	1	1	0	1	1	0	1	1	1	1	1	1	Malaria
1	1	0	1	1	0	0	1	1	1	1	0	1	1	Malaria
0	0	1	0	1	0	1	0	0	1	1	0	0	0	Yellow Fever
0	0	0	1	0	1	1	1	1	0	1	1	1	1	Malaria
0	1	0	0	1	1	0	1	0	1	1	1	1	1	Malaria
1	1	0	1	1	1	1	0	1	0	0	1	0	1	Malaria
1	1	0	1	1	1	1	1	1	1	1	1	0	1	Malaria
1	0	0	1	1	0	0	0	0	1	1	0	0	0	Zika
1	0	1	0	0	0	1	0	1	1	0	0	1	0	JE
1	1	0	1	0	0	1	1	1	1	1	1	0	1	Malaria
0	0	0	1	1	0	1	1	1	0	1	1	0	1	Plague
0	1	0	0	0	1	0	0	0	0	0	0	1	0	Tungiasis
0	1	1	0	0	1	1	0	1	0	1	1	1	1	Malaria
1	1	0	1	1	1	0	1	1	0	0	1	1	0	Malaria
1	0	1	1	0	1	1	1	1	1	1	1	1	1	Malaria
1	1	1	0	1	0	0	0	0	0	1	1	1	1	Malaria
1	1	1	1	1	1	0	0	0	0	1	0	0	1	Dengue