A Computational Approach for Medical Diagnosis Via the Decision Tree-Based Machine Learning Algorithms: A Case Study of Diagnosis Breast Cancer.

A THESIS SUBMITTED IN (PARTIAL) FULFILLMENT OF THE
REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE
APPLIED MATHEMATICS
DEPARTMENT OF MATHEMATICS, FACULTY OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
YEAR 2019
KMITL-2019-SC-M-001-002

การคำนวณเพื่อการวินิจฉัยทางการแพทย์บนพื้นฐานของขั้นตอนวิธีการเรียนรู้ของเครื่องด้วยวิธีต้นไม้การตัดสิน: กรณีศึกษาการวินิจฉัยมะเร็งเต้านม

A Computational Approach for Medical Diagnosis Via the Decision Tree-Based Machine Learning Algorithms: A Case Study of Diagnosis Breast Cancer.

จิราภรณ์ เจริญพงษ์

JIRAPHORN CHAROENPONG

หัวข้อวิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาคณิตศาสตร์ประยุกต์
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ.2562
KMITL-2019-SC-M-001-002

A Computational Approach for Medical Diagnosis Via the Decision Tree-Based Machine Learning Algorithms: A Case Study of Diagnosis Breast Cancer.

JIRAPHORN CHAROENPONG

A THESIS SUBMITTED IN (PARTIAL) FULFILLMENT OF THE
REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE
APPLIED MATHEMATICS
DEPARTMENT OF MATHEMATICS, FACULTY OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
YEAR 2019

KMITL-2019-SC-M-001-002

หัวข้อวิทยานิพนธ์	การคำนวณเพื่อการวินิจฉัยทางการแพทย์บนพื้นฐานของขั้นตอนวิธีการเรียนรู้ของเครื่องด้วยวิธีต้นไม้การตัดสิน: กรณีศึกษาการวินิจฉัยมะเร็งเต้านม
ชื่อนักศึกษา	นางสาว จิราภรณ์ เจริญพงษ์
รหัสประจำตัว	60605091
ปริญญา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาคณิตศาสตร์ประยุกต์
ภาควิชา	คณิตศาสตร์
คณะ	วิทยาศาสตร์
มหาวิทยาลัย	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.)
พ.ศ.	2562
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผศ.ดร.บุษยมาส พิมพ์พรรณชาติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	รศ.ดร.วรรณพงษ์ เตรียมโพธิ์

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์คือการพัฒนาแบบจำลองต้นไม้การตัดสินใจเพื่อวินิจฉัยมะเร็งเต้านมให้มีความถูกต้องในการทำนายมากขึ้น ข้อมูลที่ใช้ในการศึกษาคือข้อมูลของผู้ที่เข้ารับการตรวจวินิจฉัยมะเร็งเต้านมจำนวน 699 คน และปัจจัยในการวินิจฉัย 9 ปัจจัย จาก UCI Machine Learning DataSet ผลการศึกษาพบว่าเมื่อนำข้อมูลมาสร้างแบบจำลองโดยใช้ต้นไม้ตัดสินใจมีความถูกต้องในการทำนาย 95.5 % โดยงานวิจัยนี้ใช้วิธีการแบ่งแยกร้อยละเพื่อแบ่งข้อมูลเป็นสองส่วนคือ ชุดฝึกฝนข้อมูลและชุดทดสอบข้อมูล เมื่อใช้วิธีดังกล่าวพบว่าแบบจำลองที่ให้ค่าความถูกต้องมากที่สุดคือ 95.4% จากนั้นเราจึงนำแบบจำลองที่ใช้ชุดฝึกฝนข้อมูล 95% มาทดสอบด้วยข้อมูล 100% พบว่าแบบจำลองดังกล่าวให้ค่าความถูกต้องเป็น 98.86% ซึ่งค่าความถูกต้องเพิ่มขึ้นจากเดิม 3.36% ถ้าหากวิธีต้นไม้การตัดสินใจมีความน่าเชื่อถือ และถ้าผลลัพธ์ของมันเป็นที่ยอมรับได้ก็อาจทำให้วิธีการนี้สามารถสร้างประโยชน์ในทางการแพทย์ได้ ผลลัพธ์ของงานวิจัยนี้พบว่าการตรวจเพียง 8 ปัจจัยก็เพียงพอต่อการวินิจฉัยมะเร็งเต้านม โดยประโยชน์ของผลลัพธ์นี้อาจช่วยในการลดงบประมาณของโรงพยาบาลในการตัดปัจจัยที่ไม่สำคัญออกไป

คำสำคัญ : การทำเหมืองข้อมูล การแบ่งร้อยละ การวินิจฉัย ต้นไม้การตัดสินใจ มะเร็งเต้านม

Thesis Title	A Computational Approach for Medical Diagnosis Via the Decision Tree-Based Machine Learning Algorithms: A Case Study of Diagnosis Breast Cancer.
Student Name	Miss Jiraphorn Charoenpong
Student ID	60605091
Degree	Master of Science (Applied Mathematics)
Department	Mathematics
Faculty	Science
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)
Year	2019
Thesis Advisor	Asst. Prof. Dr.Busayamas Pimpunchat
Thesis Co-Advisor	Assoc. Prof. Dr.Wannapong Triampo

Abstract

This research aims to develop the model for more accurate diagnosis based on the decision tree model and find the performance of the decision tree model that gives the best accuracy. The objective of this research is to develop the decision tree model for diagnosing breast cancer, by using data from the UCI Machine Learning DataSet from Wisconsin Repository. We found that the data 699 instances to create a decision tree without the Percentage method has the accuracy to 95.5%. We use the percentage split to develop the model. It found that the model given the highest accuracy is divided the dataset as 95% or 664 instances for creating the model and 10% or 35 instances for testing the model, which this model has 95.4% of the accuracy. Then, we brought the 699 instances to test decision trees model and provide the highest accuracy with a percentage split. The results found 98.86% of the accuracy, it increased to 3.36%. Our result, if this algorithm is more acceptable and reliable. The method can also be used in medicine and we hope to be able to help to support the decision of the physician and to help patients get faster treatment. And we hope that it can help save the budget in diagnosis because unnecessary factor has cut from our study. In this research, we found the factors no need to check for all the 9 factors but use only 8 factors were able to diagnose the disease as well, which can save budget and reduce the time to diagnosis.

Keywords : Breast Cancer, Data mining, Decision tree, Diagnosis, Percentage Split.

Acknowledgements

I would like to acknowledge and thank the following important people we have supported me. Firstly, I would like to express my gratitude to my advisors, Asst. Prof. Dr.Busayamas Pimpunchat and Assoc. Prof. Dr.Wannapong Triampo for unwavering support, guidance and insight throughout this thesis. Finally, I would like to thank all my close-friends and family. They all encouraged and believed in me. They have all help me to focus on what has been a hugely rewarding and enriching process.

Jiraphorn Charoenpong

Table of Contents

List of Tables

Table **Page**

List of Figures

Figure **Page**

Chapter 1

Introduction

In this research, we will discuss the research motivation by adapting some computations of the thyroid mathematical model, the objective of the study, the scope of the study and the benefits of the study.

1.1 Research Motivation

Breast cancer is the top cancer in the developing world(Global Health Estimates, WHO 2015). For women, the 3 most commonly diagnosed cancers are breast, lung, bronchus, and colorectum, which collectively represent one-half of all cases; breast cancer alone is expected to account for 30% of all new cancer diagnoses in women [1].

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data [2]. The decision tree is a technique in data mining, which is a method of data classification. By using this technique to diagnose breast cancer and validate the decision tree model to be able to use it. In the past several years, a decision tree was applied to diagnostics, such as Shouman M. et al. using J4.8 Algorithm Decision Tree to the diagnosis of heart disease by using other discretization techniques [3]. Sumbaly R. et al. using decision tree for diagnosis of Breast Cancer by using the J4.8 algorithm in WEKA, the results of this study are the error rates and accuracy as 5.436%, 94.564% , respectively [5]. There is also a comparison of performance by dividing the data into two parts, as a training and a testing dataset, which contain several based such as cross-validation and percentage split. Muntham D. et al. was applied the Diagnosis of the Respiratory System by using the Decision Tree algorithm i.e. ID3, C4.5, and CART and the validity of the decision tree with cross-validation and percentage split method [6]. So, we choose this method in the research.

The aim of this paper is to present the decision tree model for breast cancer. By using the decision tree method to create the model. The advantage of this method is suitable for large data. To develop a model more accuracy using percentage split. We are using the measure which included accuracy, sensitivity, and specificity for investigating the performance of the model and choose the model that gives the best performance.

1.2 Objectives of the study

1) To studying Decision Tree to diagnose breast cancer.

2) To develop a model more accuracy by using the percentage split method.

1.3 Scopes of the study

1) To study Decision Tree and percentage split method.

2) To study the diagnosis of breast cancer by using the DataSet UCI from Wisconsin.

3) To compare the accuracy of the result in research with accuracy in other research.

1.4 Research methodology

1) Problem Identification

2) Studying on Decision Tree and percentage split method.

3) Studying on breast cancer data.

4) Researching about the Decision Tree for diagnosis.

5) Applying the Decision Tree with the breast cancer DataSet.

6) To analyze the result from the Decision Tree model.

7) Modifying the accuracy using percentage split.

8) Summarizing the results.

Table 1.1: (Research methodology)

Activity	Time frame				
	2018			2019	
	Apr. - Jun.	Jul. - Sep.	Oct. - Dec.	Jan. - Mar.	Apr. - Aug.
Step 1					
Step 2					
Step 3					
Step 4					
Step 5					
Step 6					
Step 7					

1.5 Benefits of the study

1) The result of this study be useful to medical for decision support of a physician.

2) To help patients get faster diagnosis of the treatment

3) The result of this study could help to save the budget due to the parameter of diagnosis reduce.

Chapter 2

Methodologies and Literature reviews

2.1 Definitions

2.1.1 Decision Tree

Decision Tree is a tree similar hierarchical structure that compose of branches and three types of root node, internal node and leaf node respectively that correspond to the sequence of decision rules [8]. Figure 2.1 show the component of decision tree. There are many types of Decision Trees. The dissimilarity between them is the mathematical model that is used in choosing the splitting attribute in extracting the Decision Tree rules. The research tests the three most generally used types: Information Gain, Gini Index, and Gain Ratio. Different decision tree algorithms are ready to use to classify the data, such as Gini index in CART, information gain in ID3 and gain ratio C4.5 [9].

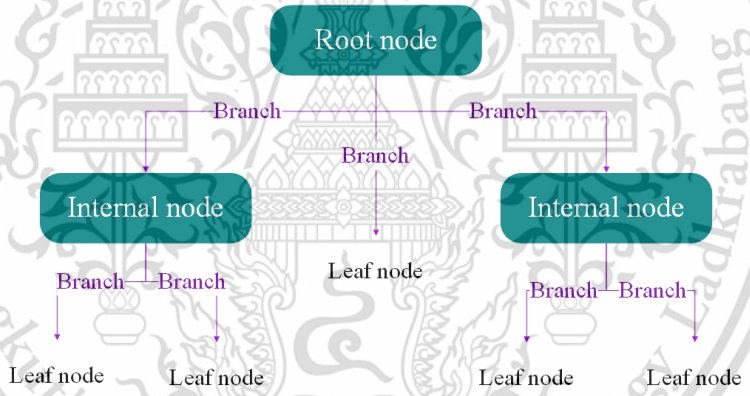


Figure 2.1: Basic Structure of Decision tree

Let the data partition D be a training set of class-labeled tuples, Suppose the class label attribute has m distinct values defining m distinct classes C_i for $i = 1, 2, ..., m$. Let $C_{i,D}$ be the set of tuples of class C_i in D. Let D and C_i, D denotes the number of tuples in $|D|$ and $|C_{i,D}|$, respectively. The ID3 technique to build a decision tree is based on information theory and attempts to minimize the expected number of comparisons. Entropy is used to measure the amount of uncertainty in the data set. Defined by the formula (2.1) [3, 4, 10].

$$Entropy\,(D) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.1}$$

where p_i is the probability that an arbitrary tuple in D belongs to class C_i.

These conditions are:

1. $S(p_1, p_2, ..., p_n)$ is a continuous function.

2. $f(n) = S(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$ is a monotonically increasing function of n.
3. Composition law for compound experiments: $S(AB) = S(A) + \sum_{k=1}^{m} p_k S(B|A)$

Consider an experiment in which we randomly pick 1 object out of N objects. The probability of picking any object is $\frac{1}{N}$. The uncertainty of this experiment is

$$f(N) = S(\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}) \tag{2.2}$$

Discrimination the N objects into m groups. Each group k contains n_k objects and k runs from 1 to m, $\sum_{k=1}^{m} n_k = N$. In the first step, resampling one of the m groups, the probability of picking group k is

$$p_k = \frac{n_k}{N} \tag{2.3}$$

Suppose group k is selected in the first step, then the probability of selecting one object in the second step is

$$\frac{1}{n_k} \tag{2.4}$$

The expected value for the uncertainty in the second step is

$$\sum_{k=1}^{m} p_k f(n_k) \tag{2.5}$$

Hence

$$f(N) = S(p_1, \ldots, p_m) + \sum_{k=1}^{m} p_k f(n_k) \tag{2.6}$$

A special case of $n_1 = n_2 = n_3 = \cdots = n$, $p_k = \frac{1}{m}$ for all k. Every group has n objects, $n \cdot m = N$

$$\begin{aligned} \sum_{k=1}^{m} p_k f(n_k) &= p_1 f(n_1) + p_2 f(n_2) + ... + p_m f(n_m) \\ &= p_1 f(n) + p_2 f(n) + ... + p_m f(n) \\ &= f(n) \cdot (p_1 + p_2 + ... + p_m) \\ &= f(n) \cdot m(\frac{1}{m}) = f(n) \end{aligned} \tag{2.7}$$

Then,

$$f(N) = S\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) + f(n) \tag{2.8}$$

$$f(mn) = f(m) + f(n) \tag{2.9}$$

Thus

$$f(m) = log(m) \tag{2.10}$$

In general case

$$
\begin{aligned}
logN &= S(p_1, \ldots, p_m) + \sum_{k=1}^{m} p_k \ log(n_k) \\
S(p_1, \ldots, p_m) &= logN - \sum_{k=1}^{m} p_k \ log(n_k) \\
S(p_1, \ldots, p_m) &= -\left(-\sum_{k=1}^{m} p_k logN + \sum_{k=1}^{m} p_k \ log(n_k) \right) \\
&= -\sum_{k=1}^{m} p_k log\frac{n_k}{N} \\
&= -\sum_{k=1}^{m} p_k log p_k
\end{aligned}
\tag{2.11}
$$

However, it is quite likely that the partitions will be impure. Since a partition may contain a collection of tuples from different classes rather than from a single class, so must be computed $Average(Entropy(D))$ which is the amount of data that is used to divide the data set D is a subset. This amount is measured by (2.12)[3, 4, 10]

$$
Average(Entropy(D)) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} \times Entropy(D) \tag{2.12}
$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the jth partition. $Average(Entropy(D))$ is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information required, the greater the purity of the partitions.

Information gain is defined as the difference between the original information requirement and the new requirement. After that is (2.13)[2, 3, 4, 10],

$$
InformationGain(D) = Entropy(D) - Average(Entropy(D)) \tag{2.13}
$$

It tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest Entropy, Information Gain is chosen as the splitting attribute at node N.

The C4.5 algorithm [9], a successor of ID3, uses an extension to Entropy known as a gain ratio, which attempts to overcome this bias. It applies a kind of normalization to Entropy using a "split information" value defined analogously with (2.11) as (2.14) [2, 3, 4, 10]

$$
SplitInfo_A(D) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} \times log_2(\frac{|D_j|}{|D|}) \tag{2.14}
$$

This value represents the potential information generated by splitting the training data set D into v partitions, corresponding to the v outcomes of a test on attribute A.

It differs from information gain, which measures the information with respect to the classification that is acquired based on the same partitioning. The gain ratio is

defined as (2.5) [4, 7]. The attribute with the maximum gain ratio is selected as the splitting attribute.

$$GainRatio(D) = \frac{InformationGain(A)}{SplitInfo_A(D)} \tag{2.15}$$

Decision Tree: Example

In this part, we're giving an example using Decision Tree for decides golf whether or not to play each day on the basis of the weather. The data consists of 4 factors: outlooks (weather: sunny, overcast, rainy), Temp. (Temperature), Humidity and Windy. Table 2.1 shows data of 14 days of weather observation related to decision golf play. [4]

Table 2.1: Data for the Golf Example.

outlook	Temp.(F)	Humidity(%)	Windy	Class
sunny	75	70	true	play
sunny	80	90	true	don't play
sunny	85	85	false	don't play
sunny	72	95	false	don't play
sunny	69	70	false	play
overcast	72	90	true	play
overcast	83	78	false	play
overcast	64	65	true	play
overcast	81	75	false	play
rain	71	80	true	don't play
rain	65	70	true	don't play
rain	75	80	false	play
rain	68	80	false	play
rain	70	96	false	play

The creation of the decision tree model will select the attributes associated with the class as the root node by using Information Gain (IG). This value is calculated from the following equation:

$$\begin{aligned} Entropy(parent) &= -p(play) \times log_2\,(play) - p(don'tplay) \times log_2\,(don'tplay) \\ &= \frac{-9}{14} \times log_2 \frac{9}{14} - \frac{5}{14} \times log_2\,(\frac{5}{14}) \\ &= 0.97 \end{aligned}$$

We calculate the entropy for the Attributes Outlook:

$$
\begin{aligned}
Entropy(Ouklook = sunny) &= -p(play) \times log_2\,(play) - p(don'tplay) \times log_2\,(don'tplay) \\
&= \frac{-2}{5} \times log_2 \frac{2}{5} - \frac{3}{5} \times log_2\,(\frac{3}{5}) \\
&= 0.97
\end{aligned}
$$

$$
\begin{aligned}
Entropy(Ouklook = rain) &= -p(play) \times log_2\,(play) - p(don'tplay) \times log_2\,(don'tplay) \\
&= \frac{-3}{5} \times log_2 \frac{3}{5} - \frac{-2}{5} \times log_2\,(\frac{2}{5}) \\
&= 0.97
\end{aligned}
$$

$$
\begin{aligned}
Average(Entropy(outlook)) &= p(outlook = sunny) \times Entropy(outlook = sunny) \\
&\quad + p(outlook = overcast) \times Entropy(outlook = overcast) \\
&\quad + p(outlook = rain) \times Entropy(outlook = rain) \\
&= (\frac{5}{14} \times 0.97) + (\frac{4}{14} \times 0) + (\frac{5}{14} \times 0.97) \\
&= 0.68
\end{aligned}
$$

$$
\begin{aligned}
InformationGain(outlook) &= Entropy(parent) - Average(Entropy(outlook)) \\
&= 0.94 - 0.68 \\
&= 0.24
\end{aligned}
$$

$$
\begin{aligned}
InformationGain(Temperature) &= Entropy(parent) - Average(Entropy(Temperature)) \\
&= 0.94 - 0.91 \\
&= 0.03
\end{aligned}
$$

$$
\begin{aligned}
InformationGain(Humidity) &= Entropy(parent) - Average(Entropy(Humidity)) \\
&= 0.94 - 0.79 \\
&= 0.15
\end{aligned}
$$

$$
\begin{aligned}
InformationGain(Windy) &= Entropy(parent) - Average(Entropy(Windy)) \\
&= 0.94 - 0.89 \\
&= 0.05
\end{aligned}
$$

The outlook is the higher information gain, then outlook will be the roof of tree. Figure 2.2 Show the decision tree model that the root is outlook. Next, we find the information gain for the rain branch.

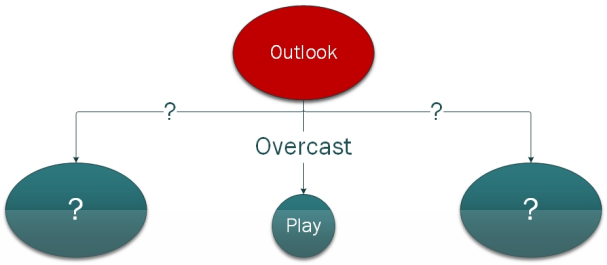


Figure 2.2: Decision tree model that the root is outlook.

$$Entropy(parent) = -p(play) \times log_2\,(play) - p(don'tplay) \times log_2\,(don'tplay)$$
$$= \frac{-3}{5} \times log_2 \frac{3}{5} - \frac{-2}{5} \times log_2\,(\frac{2}{5})$$
$$= 0.97$$

$$InformationGain(Temperature) = Entropy(parent) - Average(Entropy(Temperature))$$
$$= 0.97 - 0.95$$
$$= 0.02$$

$$InformationGain(Humidity) = Entropy(parent) - Average(Entropy(Humidity))$$
$$= 0.97 - 0.95$$
$$= 0.02$$

$$InformationGain(Windy) = Entropy(parent) - Average(Entropy(Windy))$$
$$= 0.97 - 0$$
$$= 0.97$$

Windy is the higher information gain, then it will be the internal node of rain branch. we find the information gain for the sunny branch.

$$Entropy(parent) = -p(play) \times log_2\,(play) - p(don'tplay) \times log_2\,(don'tplay)$$
$$= \frac{-2}{5} \times log_2 \frac{2}{5} - \frac{-3}{5} \times log_2\,(\frac{3}{5}) = 0.97$$

$$InformationGain(Temperature) = Entropy(parent) - Average(Entropy(Temperature))$$
$$= 0.97 - 0.4$$
$$= 0.57$$

$$InformationGain(Humidity) = Entropy(parent) - Average(Entropy(Humidity))$$
$$= 0.97 - 0$$
$$= 0.97$$

Humidity is the higher information gain, then it will be the internal node of sunny branch. Decision tree model of golf playing decisions as shown in Figure 2.3. From Figure 2.3, the model can be described as follows: if outlook is overcast or cloudy, then the golfer will decide to play golf. If outlook is sunny and Humidity more than 75 then golfers will decide not to play golf.

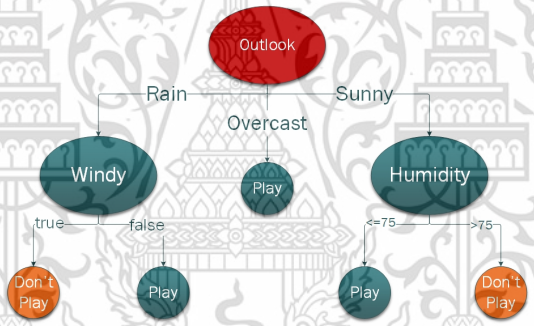


Figure 2.3: Decision Tree for the Golf Example.

2.1.2 Percentage Split (Holdout Method)

In this method, the given data are randomly partitioned into two independent sets, a training, and a test set. Typically, two-thirds of the data are allocated to the training set, and the remaining one-third is allocated for testing. The training set is used to derive the model, whose accuracy is estimated with the test set [2].

2.1.3 Measuring the Performance of a Classifier

Confusion Matrix

There are two classes, which call positive and negative, the confusion matrix consists of four cells, which can be labeled TP, FP, FN, TN as in Table 2.2 [2, 3, 10].

Where TP (True Positive): the number of positive instances that are classified as positive. TN (True Negative): the number of negative instances that are classified as negative. FP (False Positive): the number of negative instances that are classified as

Table 2.2: True and False Positives and Negatives.

Confusion Matrix	Predicted Class	
	Positive	Negative
Actual Class	TP	FP
	FN	TN

positive. FN (False Negative): the number of negative instances that are classified as negative.

2.1.4 Performance Measures

It is the task of classifying the elements of a given set into more than one group on the basis of a classification rule. There are many metrics that can be used to measure the performance of a classifier or predictor; different fields have different preferences for specific metrics due to different goals. The accuracy of a classifier is a value that indicates the ability of a measurement. Sensitivity measures the proportion of actual positives that are correctly identified as such. Specificity measures the proportion of actual negatives that are correctly identified as such. These measures are defined as follows [2, 3, 10, 11]

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.16}$$

$$Specificity = \frac{TN}{FP + TN} \tag{2.17}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.18}$$

To provide more insight. We give an example: Examples of breast cancer diagnoses, which have two alternatives, are malignant and benign. This data contains 699 patients consist of benign 458 and malignant 241 patients. From the prediction, the model can be evaluated using a confusion matrix, as shown in Table 2.3.

Table 2.3: Example of Confusion Matrix.

Confusion Matrix	Predicted Class	
	Positive	Negative
Actual Class	443(TP)	6(FP)
	15(FN)	235(TN)

From Table 2.3, it was found that 443 patients correct predicted that benign and wrong predicted of 15 patients (the patients are benign, but the model is predicted to be malignant). And, it was found that 235 patients correct predicted that malignant and

wrong predicted of 6 patients (the patients are malignant, but the model is predicted to be benign). The model can be evaluated by using the accuracy value which is approximately 97

2.2 Related Research

Machine Learning can be applied in many things such as industry, agriculture, medical, etc. Machine Learning used for the analysis of the importance of clinical parameters and their combinations for prognosis, e.g. prediction of disease progression, extraction of medical knowledge for outcome research, therapy planning and support, and the overall patient management [12]. In the past many years, there is a lot of research to use Machine Learning Algorithm to apply in cancer. In the survival analysis for cancer. This method is used to find the survival rate of cancer patients. Application of Machine Learning to Survival will help predict survival rates. This application can be developed for survival prediction in rare pathologies and have the potential to serve as the basis for the creation of decision support tools in the future [13]. At present, the software helps us to work more conveniently and save more time. And some types of software are also available for free. In the ML, there are several types of software available for use. Weka is a tool for the initial processing of data. Advantages of Weka Firstly, Weka is an open-source. Secondly, it solves the ML problem. Finally, it runs on many platforms [14]. It also allows users to select the most appropriate algorithm for the model to obtain the most accurate value [15]. Another tool that is used frequently, the R programming. R is used in machine learning for classification, regression, and clustering. There are a variety of packages in R to make ML easier to use, such as rpart [16] is a package that helps in recursive partitioning for classification, regression and survival trees. Another important package is e1701 [17]. It can be used in support vector machines, shortest path computation, bagged clustering, and Naive Bayes classifier.

Machine Learning Applications with Cancer From survey research of machine learning application with cancer. The first matter to be studied is diagnostic. A comparison of the performance of different machine learning algorithms has been studied in many researches. This research uses public information from the UCI. This data is a breast cancer test with two classes for decision making, with malignant and benign. This information is available for 699 patients, consisting of 241 of malignant and 458 of benign. The main purpose of this research is to compare and evaluate different machine learning. By using SVM, DT (C4.5), NB and k-NN to predict the diagnosis of breast cancer and comparative efficiency [18]. The results of such research showed that SVM gives the highest accuracy (97.13%) and the lowest error rate (2%) compared with other algorithms. It can be said that the SVM algorithm is a method that provides the most

accurate of this dataset for prediction diagnosis. Using ANN and LR to developing a model to predict liver cancer [19] is one of the interests by using 70% of the data to training and 30% for testing the model. The result shows that the accuracy of ANN better than the LR. In the case of survival and recurrence, there is interesting research as well. The one paper report is the prediction model of survival and recurrence for breast cancer by using machine learning [20]. The result of predict survival shows that the ANN algorithm obtained the best accuracy (83.6%) using live-one-out cross-validation as a test method. Meanwhile in the prediction of recurrence of breast cancer found that the DT algorithm has achieved the best accuracy (96.57%) by using validation like that the validation of prediction of survival. The comparison of the methods of machine learning is still many in many researches. This research uses 3 machine learning algorithms (DT (C4.5), SVM, ANN) for medical data analysis which is big data to find hidden knowledge. The data for this study is the data of breast cancer patients admitted to the Iran Center program, which has a total of 1189 patients. The results of this study showed that SVM predicted breast cancer models better than other methods with the highest correctness of 95.7% and DT with the lowest predictive value of 93.6% [21]. The next research is also a study of breast cancer. This research uses 7 machine learning algorithms consisting of NB, Trees Random Forest, 1-Nearest Neighbor, AdaBoost, SVM, RBF Network and Multilayer Perceptron. The objective of this research is to apply the seven methods of machine learning to predict the model of survival of breast cancer. The data from the study of Cancer Registry Organization of Kerman, Iran with 900 breast cancer patients, consisting of 803 living patients and 97 patients died. The results of this study showed that the Trees Random Forest predicted the best breast cancer survival model with 96% accuracy. And 1-Nearest Neighbor predicted the model to be less accurate than other methods, 91% of accuracy [22].

Shouman M. et. al. using J4.8 Algorithm Decision Tree and other discretization techniques to analyse a model of the diagnosis heart disease. The objective of this research is to show a new model that enhances the Decision Tree accuracy by apply voting. The data used in this research is heart disease data from the Cleveland Clinic Foundation. The study found that the application of Voting with the Decision Tree shows the increased accuracy [3].

Sumbaly R. et. al. using decision tree for diagnosis of Breast Cancer by using the J4.8 algorithm in WEKA. The data used in this research is Breast Cancer DataSet from the UCI Machine Learning Repository. This DataSet consist of 9 factor to diagnosis breast cancer. The results of this study are the error rates and accuracy as 5.436%, 94.564% respective. And this research also found that the diagnosis using only 5 factors are enough to diagnose breast cancer [5].

Muntham D. et. al. was applied Diagnosis of the Respiratory System by using the Decision Tree algorithm i.e. ID3, C4.5, and CART and the validity of the decision

tree with cross-validation and percentage split method. The results showed that for a patients with acute with a ratio 70:30 of the training data and the testing data, Decision Tree (C4.5) was the most effective (92.31%). For patients with acute sinusitis with ratio 70:30 of training data and testing data, Decision Tree(C4.5) was most effective (94.70%). For patients with pneumonia with ratio 50:50 of training data and testing data, the CART Algorithm was effective (94.69%) [6].

Parveen et.al studied Decision Tree to evaluate the risk of Non-Alcoholic Fatty Liver Disease in the Canadian population. The data used in this study is the risk factor of Non-Alcoholic Fatty Liver Disease from Electronic Medical Records (EMRs). The result shows that the Decision Tree method was most effective for the diagnosis of Non-Alcoholic Fatty Liver Disease risk, 93.7 of accuracy [23].

Cirkovic et.al. showed prediction model of survival and recurrence of breast cancer by using machine learning. The result of predict survival shows that the ANN algorithm obtained the best accuracy (83.6%) with using live-one-out cross-validation as a test method. Meanwhile in prediction of recurrence of breast cancer found that Decision Tree algorithm has achieved the best accuracy (96.57%) by using validation [24].

P.Hamsagayathri et.al. studied Decision Tree algorithms for Breast cancer to analyze the performance of classification. The data used in this research was study from SEER breast cancer dataset. And this research use WEKA software to breast cancer diagnosis classification. The results show that Decision Tree classifier classifies the data with 98.51% accuracy [25].

Chaurasia V. et.al used three algorithms (Naïve Bayes, Radial Basis Functions Neural Networks, J48) to develop the prediction models for breast cancer diagnosis. The data used in this research is Breast Cancer DataSet from the UCI Machine Learning Repository. This DataSet consist of 683 breast cancer cases. The results showed that Naïve Bayes is the best predictor with 97.36% accuracy. Meanwhile J48 came out the accuracy with 93.41% [26].

Chapter 3

Research Methodology

3.1 Algorithms

The algorithm to create a model for diagnosis of breast cancer by using a decision tree and choose the model that, given the best performance is shown in Figure 3.1. The steps to create a model are as follows.

1) Studying on Breast cancer DataSet from UCI.
2) Studying on classification algorithm.
3) Finding suitable methods for classifying with DataSet.
4) Studying on percentage split or holdout method
5) Dividing the data into 2 sets: training set and testing set
6) Using the training set to learning process with a decision tree to create a prediction model.
7) Using the testing set to test model.
8) Performance evaluation of breast cancer diagnosis model by using confusion matrix

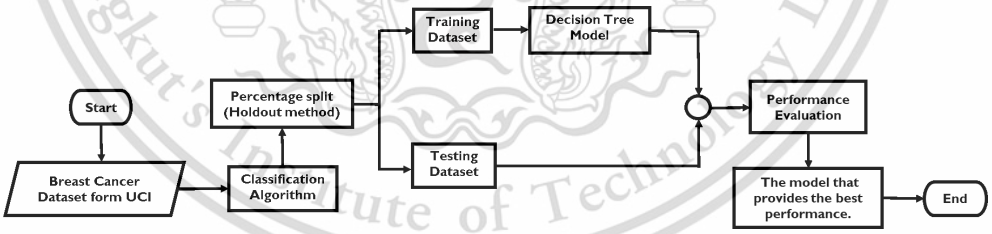


Figure 3.1: Flow Chart for Diagnosis Breast Cancer by Using the Decision Tree to Building Model with Percentage Split Method.

3.2 The root node of the Decision tree

Details of the model are shown in Figure 4.9. By the C5.0 algorithm, the attribute with the maximum gain ratio is selected as the splitting attribute [10]. Therefore, the attribute with the highest gain ratio is used as a root node of the tree.

3.3 The branch of the Decision tree

The data used to create a Decision Tree model in this paper is numeric. It consists of the following steps:

i) Sort smallest to largest,

ii) Find the midpoint between two different values, i.e. the midpoint between 1, 2 is 1.5, etc., then divide the data into two sections, divided into more than and less than or equal.

iii) Calculating Entropy of two parts, the part of less than or equal of the mid-point to be on the left side of the node and the part more than to be on the right side.

iv) Find the midpoint of all the values in the data, and then select the value to give the Gain Ratio most.

3.4 The model of the decision tree

To choose a decision tree model. We have divided the data into training Dataset and testing Dataset. The criterion for choosing a model is the accuracy of the model.From Figure 3.2 shows the accuracy of each model.

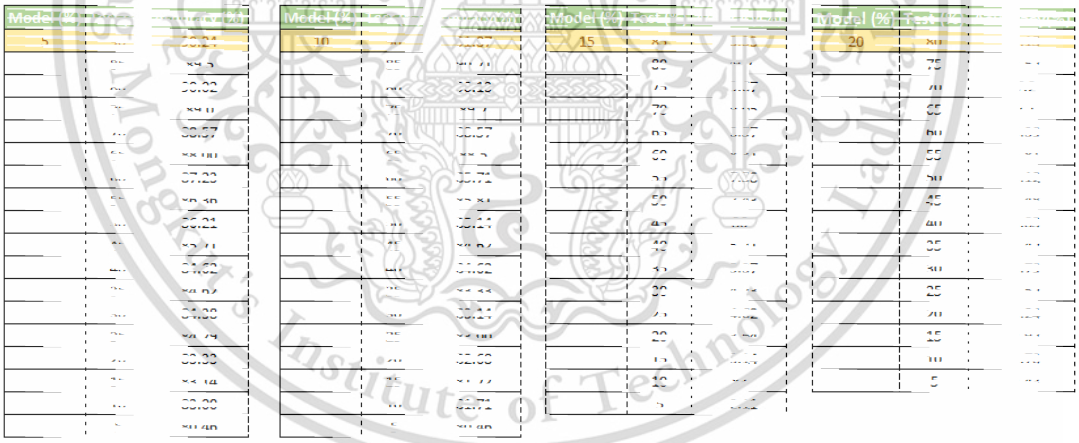


Figure 3.2: The accuracy of training Dataset (5% to 20%)

From Figures 3.2 to 3.6, it shows the accuracy of dividing data into two sets. The principle of the models using the percentage split method which is divide the data into two parts and both parts must not have duplicate data.

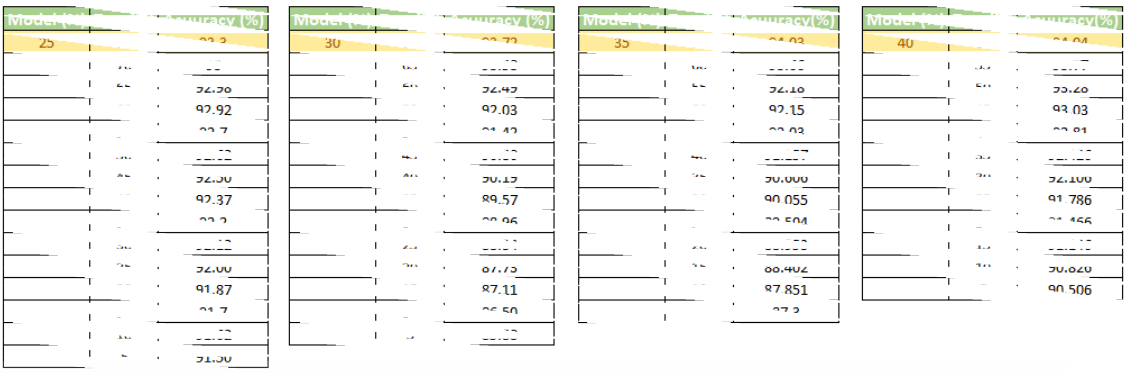


Figure 3.3: The accuracy of training Dataset (25% to 40%)

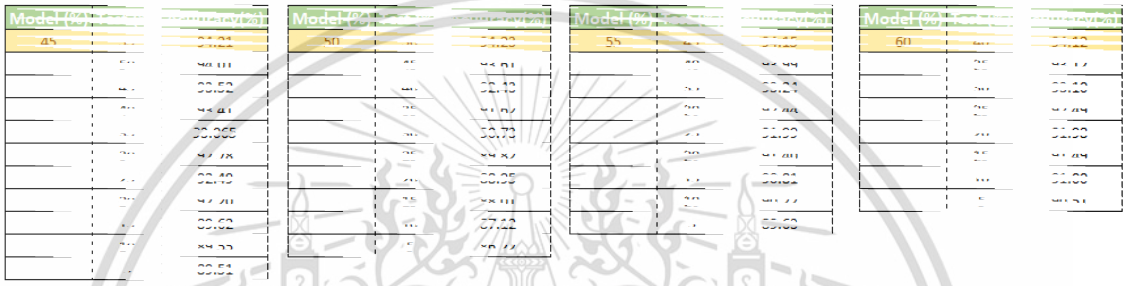


Figure 3.4: The accuracy of training Dataset (45% to 60%)

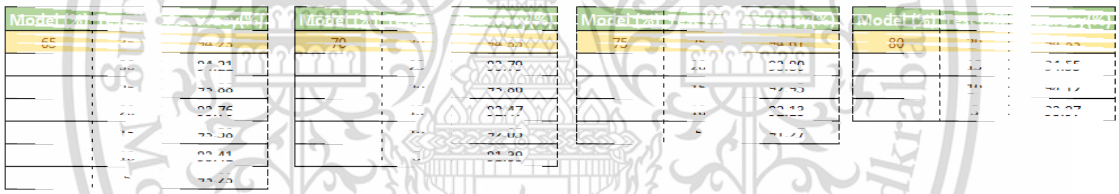


Figure 3.5: The accuracy of training Dataset (65% to 80%)

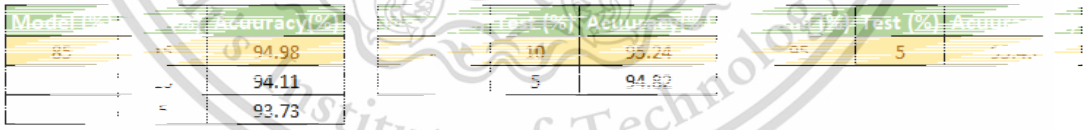


Figure 3.6: The accuracy of training Dataset (85% to 95%)

From the above figures, it shows that the test data decreases, the accuracy decreases. In other words, Training data may not cover all testing data. While the training data is the same, the test data has changed.Therefore, if the data in the test increases, it may make the model more accurate. Because there are cases that can cover the model.

Chapter 4

Results

4.1 Data Description

The data used in this study is Breast Cancer Wisconsin DataSet from UCI Machine Learning Repository [10]. This dataset having 699 instances consist of benign 458 instances and malignant 241 instances. Figure 4.1. shows the number in the attribute class of the dataset.

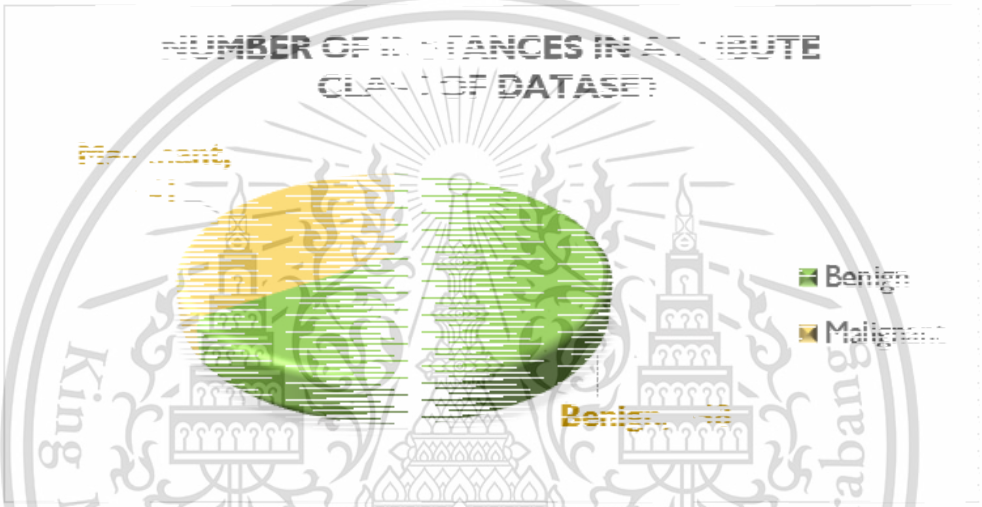


Figure 4.1: Number of Instances in Attribute Class of Dataset.

The attributes in this dataset are shown in Table 4.1. By using the attributes to diagnose, only 9 attributes because the sample code number is not important for the diagnosis of breast cancer.

4.2 To Split the data.

In this study, we used the C4.5 algorithm to model building the decision tree. We bring data 699 instances to predict the diagnosed of breast cancer with a decision tree and not yet use the Percentage split method found the accuracy is 95.5%. To develop the performance of the decision tree model, we use the Percentage Split method or Holdout method [6, 23]. The Percentage Split method by determining the default ratio is 5:95 between the training dataset and the testing dataset and add each step, 5% of the training dataset and reduce each step, 5% of the testing dataset. Since the Percentage split data is divided into 2 parts, each part is selecting data using random. Therefore, to more accurate and to the more distribution of data, we randomized data 100 times then take to make a decision tree. With this technique, it brings up more

Table 4.1: Breast Cancer Wisconsin Dataset Attributes.

No	Attribute Name	Value
1	Uniformity of Cell Shape (UoCS)	1-10
2	Uniformity of Cell Size (UoCZ)	1-10
3	Bare Nuclei (BN)	1-10
4	Bland Chromatin (BC)	1-10
5	Single Epithelial Cell Size (SECZ)	1-10
6	Normal Nucleoli (NN)	1-10
7	Marginal Adhesion (MA)	1-10
8	Clump Thickness (CT)	1-10
9	Mitoses (M)	1-10
10	Sample code number	The id number of instances
11	Class	2-Benign and 4-Malignant

accuracy, and specificity sensitivity value to find the average.

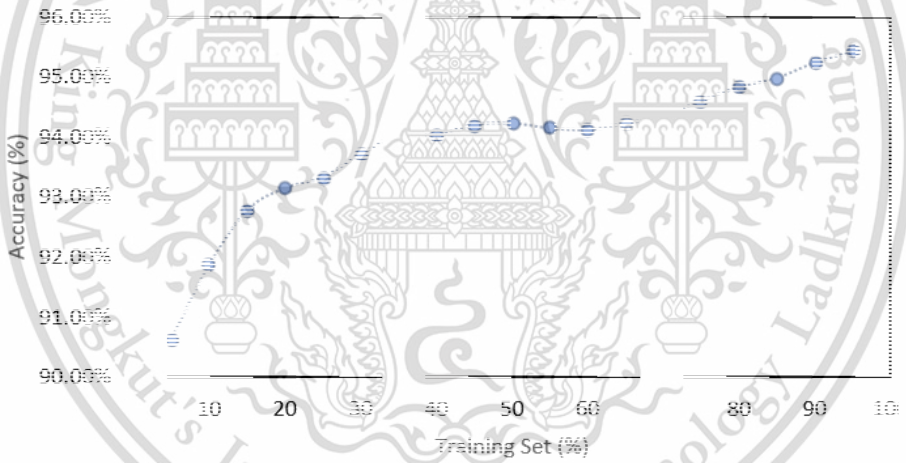


Figure 4.2: The Accuracy of 5% to 95% of the Training Dataset.

By data showing trends in Figure 4.2, it is seen that, agrees that accuracy tends to rise when the training dataset is increasing while the testing dataset is decreasing. It was found that a ratio 95:5 of the training dataset and the testing dataset which provides the highest accuracy, sensitivity, and specificity are 95.4%, 95.8 and 94.6% respectively, we conclude the information shown in Table 4.2.

4.3 To find the root node

Details of the model are shown in Figure 4.9. By the C5.0 algorithm, the attribute with the maximum gain ratio is selected as the splitting attribute [10]. Therefore, the attribute with the highest gain ratio is used as a root node of the tree.

Table 4.2: The Result of Performance Measures.

Training Data(%)	Testing Data(%)	Accuracy(%)	Sensitivity (%)	Specificity (%)
5	95	90.64	95.26	81.88
10	90	91.87	95.35	85.28
15	85	92.78	95.42	87.73
20	80	93.15	95.71	88.30
25	75	93.33	95.33	89.51
30	70	93.72	94.80	91.66
35	65	94.03	94.95	92.26
40	60	94.04	94.98	92.25
45	55	94.21	95.06	92.56
50	50	94.23	94.58	93.60
55	45	94.15	94.53	93.41
60	40	94.12	94.46	93.48
65	35	94.23	94.67	93.44
70	30	94.35	94.34	94.43
75	25	94.61	94.87	94.11
80	20	94.83	94.98	94.56
85	15	94.98	95.48	94.06
90	10	95.24	95.57	94.48
95	5	95.43	96.80	94.64

It consists of the following steps:

i) Preparing the data as the figure 4.3

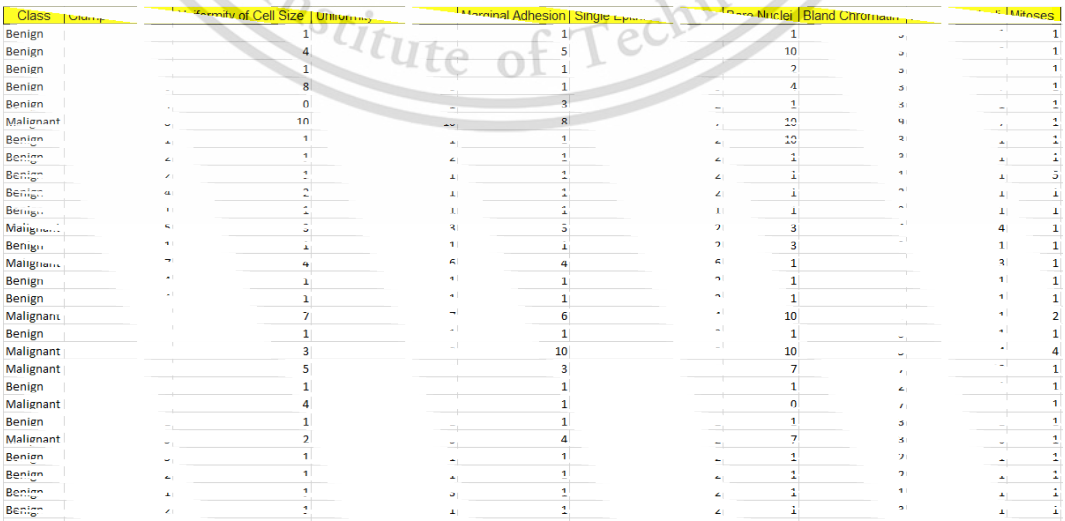


Figure 4.3: Preparing the data for find the root node.

ii) Taking the data to table for calculate as the figure 4.4

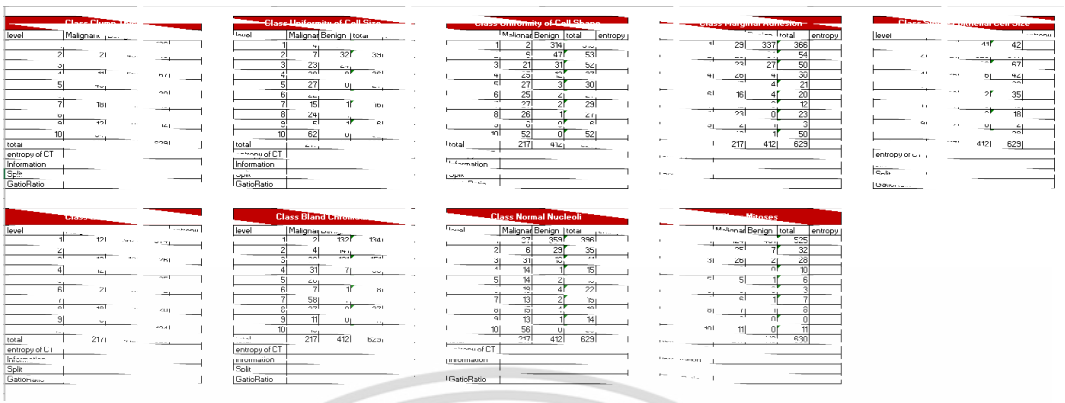


Figure 4.4: Putting the data on table for calculate the root node.

iii) Calculating the Entropy, Information gain, Split Information and Gain Ratio as figure 4.5

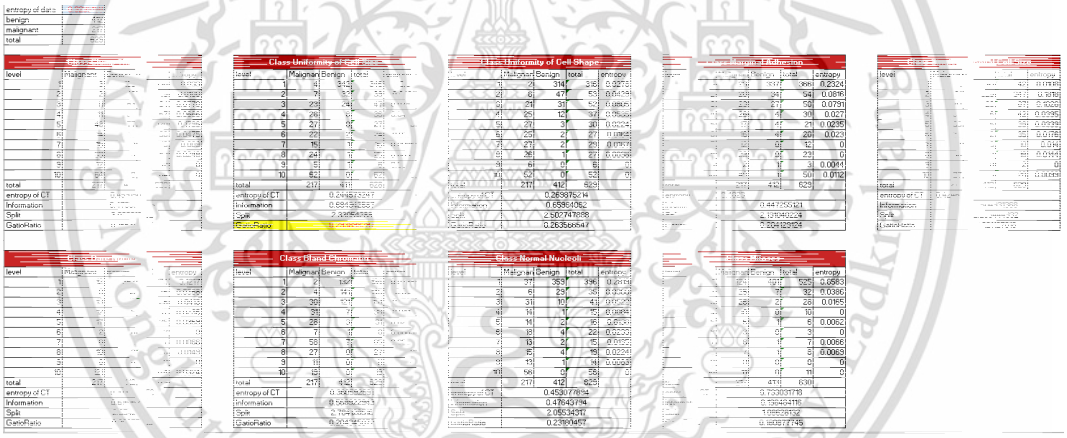


Figure 4.5: Calculating the data for find the root node.

For example, Finding the Gain Ratio of Clump Thick. Firstly, preparing the data of Uniformity of Cell Size as figure 4.6.

Next, calculating the entropy, information gain, split information and gain ration respectively. This calculation as follows : Using the training dataset which has 629 instances detect a benign of 412 instances and malignant 217 instances by calculating Entropy of 629 instances as follows

$$Entropy(629instances) = -\frac{412}{629}log_2(\frac{412}{629}) - \frac{217}{629}log_2(\frac{217}{629}) = 0.929516$$

Finding the entropy of Clump Thick as follow

$$Entropy(ClumpThick) = Entropy(Level1)+Entropy(Level2)+Entropy(Level3)+Entropy(Level4)+$$
$$Entropy(Level5)+Entropy(Level6)+Entropy(Level7)+Entropy(Level8)+Entropy(Level9)+$$
$$Entropy(Level10) = 0.033+0.0188+0.0778+0.0686+0.1796+0.0475+0.009+0.0248+0+0 =$$

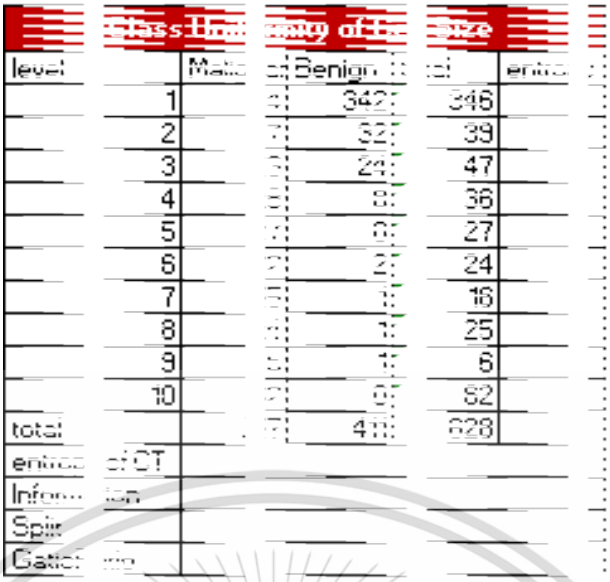


Figure 4.6: Preparing the data of Clump Thick for calculate root node

0.459150376

Then, $InformationGain(CT) = Entropy(629ins.) - Entropy(CT) = 0.47036563$

And $SplitInfo_A(ClumpThick) \approx 3.026918181.$

Hence, $GainRatio(ClumpThick) = \frac{InformationGain(ClumpThick)}{SplitInfo_A(ClumpThick)} = 0.155394176$

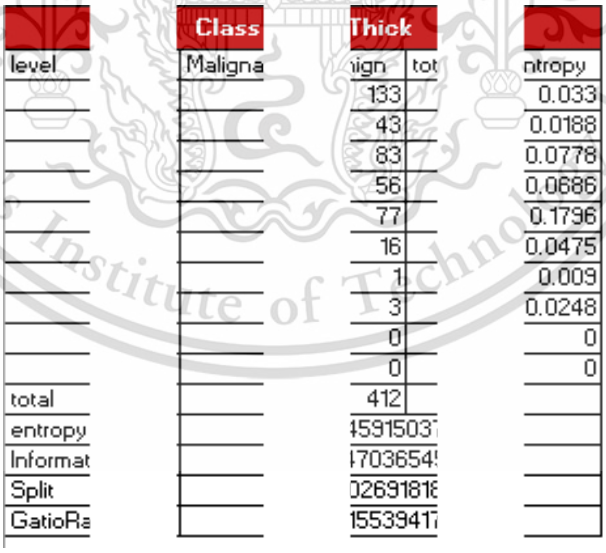


Figure 4.7: The Calculation of branch

Similarly, we can compute the Gain Ratio of other factors, in the same way, which did above. Which has the following value

Gain Ratio (Uniformity of Cell Size)= 0.293898297 , Gain Ratio (Uniformity of Cell Shape)=0.263566547, Gain Ratio(Marginal Adhesion)=0.204129124 , Gain Ratio(Bare

Nuclei)=0.20857053 , Gain Ratio(Bland Chromatin)=0.204345072 , Gain Ratio(Normal Nucleoli)=0.23180457 , Gain Ratio(Mitoses)=0.180877745 and Gain Ratio (Single Epithelial Cell Size)=0.221127018 .

Figure 4.5 shows the detail of the root node calculate. Since Gain Ratio(Uniformity of Cell Size)Given the most Gain ratio, so the root node chooses the Uniformity of Cell Size as a root.

4.4 To find the branch of tree

The data used to create a Decision Tree model in this paper is numeric. It consists of the following steps:

i) Sort smallest to largest,

ii) Find the midpoint between two different values, i.e. the midpoint between 1, 2 is 1.5, etc., then divide the data into two sections, divided into more than and less than or equal.

iii) Calculating Entropy of two parts, the part of less than or equal of the midpoint to be on the left side of the node and the part more than to be on the right side.

iv) Find the midpoint of all the values in the data, and then select the value to give the Gain Ratio most.

To get a better understanding, I'll give the example for calculating the branch of a root node, which is the Uniformity of Cell Shape. The data in the Uniformity of Cell Shape range from 1 to 10. It sorts the data from the smallest to largest and then find the midpoint between two different values.

Consequently, the midpoint in the Uniformity of Cell Size has 1.5, 2.5, 3.5, 4.5, ..., 9.5 respectively. Bring up the midpoint obtained to divide the two ranges, such as >1.5,≤ 1.5 etc. For creating a decision tree model, using the training dataset which has 629 instances detect a benign of 412 instances and malignant 217 instances by calculating Entropy of 629 instances as follows

$$Entropy(629instances) = -\frac{412}{629}log_2(\frac{412}{629}) - \frac{217}{629}log_2(\frac{217}{629}) = 0.929516$$

In the section of ≤ 1.5 detects benign 314 instances, malignant 2 instances and in the section of >1.5 detects benign 98 instances, malignant 215 instances. To find the Gain Ratio can be calculated as follows

$$Entropy(\leq 1.5) = \frac{316}{629} \times \left(-\frac{314}{316}log_2(\frac{314}{316}) - \frac{2}{316}log_2(\frac{2}{316})\right)$$

$$Entropy(> 1.5) = \frac{313}{629} \times \left(-\frac{98}{313}log_2(\frac{98}{313}) - \frac{215}{313}log_2(\frac{215}{313})\right)$$

So, $Entropy(1.5) = 0.027796 + 0.446221 = 0.474017$

Then, $InformationGain(1.5) = Entropy(629ins.) - Entropy(1.5) = 0.455499.$

And $SplitInfo_A(1.5) \approx 0.999984.$

Hence, $GainRatio(1.5) = \frac{InformationGain(1.5)}{SplitInfo_A(1.5)} = 0.507076$

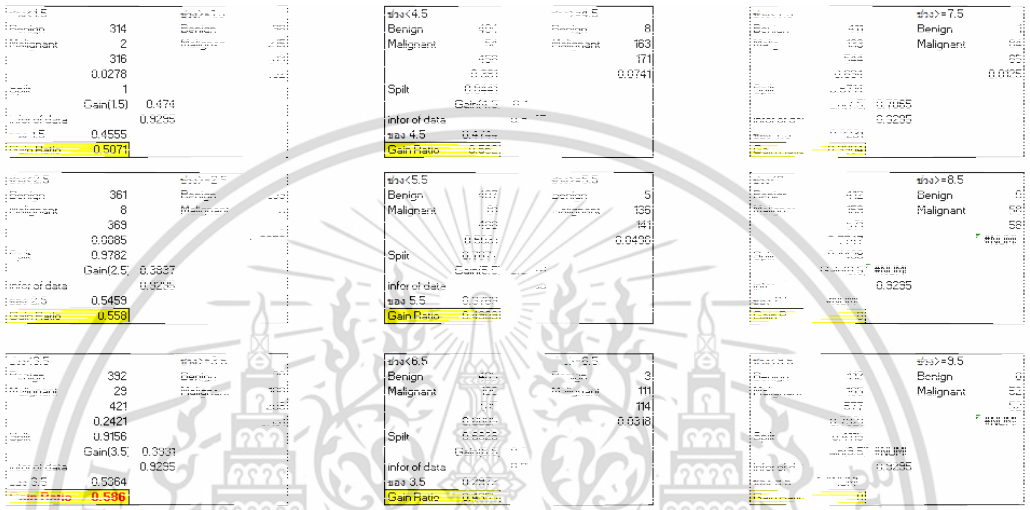


Figure 4.8: The Calculation of branch

Similarly, we can compute the Gain Ratio of other values, in the same way, which did above. Which has the following value Gain Ratio (2.5)= 0.558009 , Gain Ratio (3.5)=0.585805 ,Gain Ratio(4.5)=0.561974, Gain Ratio(5.5)=0.490921,Gain Ratio(6.5)=0.435272 ,Gain Ratio(7.5)=0.390407,Gain Ratio(8.5)=0 and Gain Ratio (9.5)=0. Figure 4.3 shows the detail of the branch of tree calculate. Since Gain Ratio(3.5)Given the most Gain ratio, so the root node chooses 3.5 as a branch.

4.5 Results of decision tree

From Figure 4.9 the decision tree model can be converted to classification rules, which have 13 rules. When a biopsy was found Uniformity of Cell Size > 3.5, Uniformity of Cell Shape > 1.5 and Bland Chromatin > 1.5 have the opportunity to malignant. But Uniformity of Cell Shape > 3.5, Uniformity of Cell Size ≤ 1.5 have the opportunity to benign, etc., it is summarized in the following Table 4.3.

Decision tree model in Figure 4.9 provides an important factor used to diagnose, namely Uniformity of Cell Shape and other factors are Uniformity of Cell Size, Bland Chromatin, Single Epithelial Cell Size, Normal Nucleoli, Clump Thickness, Marginal Adhesion, and Mitoses. In the decision tree model, we are not found the Bare Nuclei. it

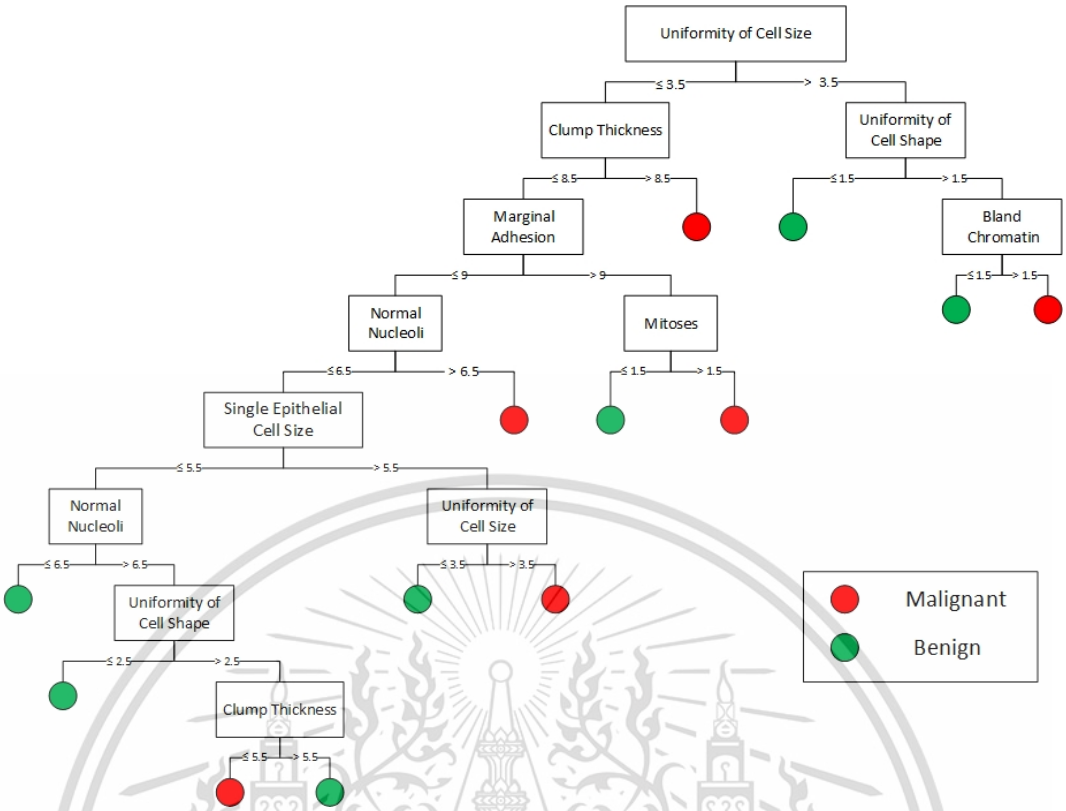


Figure 4.9: Decision tree model for testing data set by using percentage split 95:5.

means this factor impact little or none the diagnosis of breast cancer. The details of the decision tree model for the diagnosis of breast cancer are shown in Table 4.

From the confusion matrix of this model, as shown in Table 4.4, it is found that the number of the correctly classified instance, as 33 and the number of the wrongly classified instance as 2. Here TP is the number of the benign instance predicted as a benign instance. FN is the number of the benign instance predicted as a malignant instance. FP is the number of the malignant instance predicted as a benign instance. And TN is the number of the malignant instance predicted as a malignant.

It is noted that the false negative is the value of prediction error. It indicates the patients who have malignant but are predicted as benign. It is considered as the dangerous misdiagnosed which could affect the patient greatly. Also, it may cause the patient to die from a faulty diagnosis. In the data of true positive of benign, we found the value of a factor have not higher than 5. A true negative or classes of malignant of each factor are high more than 5 and less than 5 which found very few. But the false-negative appears that some factor is very high and some factors are low which may lead to a wrong prediction because each factor is not going in the same direction. Therefore, if the results of the diagnosis of the patients have the characteristic factors as the above, then the patients should be diagnosed carefully checked again. The model is developed in this paper. Therefore, taking the data for 699 instances to testing this model by the prediction from the confusion matrix shown in Table 4.4.

Table 4.3: The Result of Decision Tree Model for the Diagnosis Breast Cancer.

No.	UoCZ.	UoCS.	BC.	SECZ.	NN.	MA.	CT.	M.	Diagnosis.
1	≤ 3.5						>8.5		Malignant
2	≤ 3.5						≤ 8.5	>1.5	Malignant
3	≤ 3.5						≤ 8.5	≤ 1.5	Benign
4	≤ 3.5				>9	>6.5	≤ 8.5		Malignant
5	≤ 3.5			>5.5	≤ 9	>6.5	≤ 8.5		Malignant
6	All level			>5.5	≤ 9	≤ 6.5	≤ 8.5		Benign
7	≤ 3.5			≤ 5.5	≤ 3.5	≤ 6.5	≤ 8.5		Benign
8	≤ 3.5			≤ 5.5	3.5 to 9	≤ 6.5	≤ 8.5		Benign
9	≤ 3.5	> 2.5		≤ 5.5	3.5 to 9	≤ 6.5	≤ 5.5		Malignant
10	≤ 3.5	≤ 2.5		≤ 5.5	3.5 to 9	≤ 6.5	5.5 to 8.5		Benign
11	> 3.5	≤ 1.5	≤ 1.5				>6.5		Malignant
12	> 3.5	≤ 1.5	≤ 1.5				≤ 6.5		Benign
13	> 3.5	≤ 1.5	≤ 1.5				≤ 6.5		Benign

Table 4.4: Confusion Matrix of Decision Tree for Diagnosis of Breast Cancer by using 35 instances.

Confusion Matrix	Predicted Class		Total
	Malignant	Benign	
Pred.Malignant	21(TP)	1(FP)	22
Pred.Benign	1(FN)	12(TN)	13
Total	22	13	35

From table 4.5, the accuracy, sensitivity, and specificity are 97.0%, 96.7%, and 97.5%, respectively. We found that the rate of accuracy increases to 1.5% when using the Percentage Split method. Therefore it is reasonable to say that the method in this research, provides more accurate than that of the J48 algorithm method in [5].

Table 4.5: Confusion Matrix of Decision Tree for Diagnosis of Breast Cancer by using 699 instances.

Confusion Matrix	Predicted Class		Total
	Malignant	Benign	
Pred.Malignant	443(TP)	6(FP)	447
Pred.Benign	15(FN)	235(TN)	252

Chapter 5

Conclusion and Discussion

This research aims to develop the model for a more accurate diagnosis based on the decision tree model and find the performance of the decision tree model that gives the best accuracy. We focused on Breast Cancer by using data from the UCI Machine Learning Data Set form Wisconsin Repository. We used a decision tree which is one of data mining technique and the Percentage Split for classifying data into the training dataset and testing dataset. We applied the C5.0 Algorithm for developing the model. The model provides the most accurate scheme for 669 instances for machine learning and 35 samples for testing a decision tree model. Having used the random data 100 times, we then got the average of accuracy, sensitivity, and specificity as 95.4%, 95.8%, and 94.6%, respectively. If such a method is acceptable and can be trusted, it can be applied in medical applications. We also hope this model would help support the decision of the physician resulting in helping patients get faster treatment. Meanwhile, applying our developed model to other data sets for model developing and optimization is undergoing process.

References

[1] RL. Siegel, KD. Miller, A. Jemal. Cancer Statistics, 2017. CA Cancer J Clin., 7-30, 2017.

[2] J. Han, M. Kamber, J. Pei, Data mining, Concepts and Techniques 3rd ed., USA, 2011.

[3] M.Shouman, T. Turner, R. Stocker, Using Decision Tree for Diagnosing Heart Disease Patients. Proceedings of the 9th Australasian Data Mining Conference, Ballarat, Australia, 2011.

[4] M. Bramer, Principles of Data Mining (Third Edition), Undergraduate Topics in Computer Science, Springer Verlag London Ltd., 2016.

[5] R. Sumbaly, N. Vishusri, S. Jeyalatha, Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. International Journal of Computer Applications Volume 98-No.10, 16-24, 2014.

[6] D. Muntham, L. Ingsrisawang, An Application of Decision Tree Algorithms for Diagnosis of the Respirator System: A Case Study of Pranakorn Sri Ayudthaya Hospital. Journal of Health Systems Research Vol.4 No.1, 72-81, 2010.

[7] S. Auwatanamongkol. Data Mining. Bangkok Block Ltd., Part, Bangkok, 31-41, 2014.

[8] S. Dharm, Naveen C.and Jully S., Analysis of Data Mining Classification with Decision tree Technique, Global Journal of Computer Science and Technology Software & Data Engineering, Vol. 13 Issue 13, 2013.

[9] A. N. Richter, & Khoshgoftaar, T. M., A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artificial Intelligence in Medicine., 2018.

[10] เอื้อวัฒนามงคล สุรพงศ์, การทำเหมืองข้อมูล, สำนักพิมพ์สถาบันฑิตพัฒน์บริหารศาสตร์, 2010.

[11] สินสมบูรณ์ทอง สายชล, การทำเหมืองข้อมูล, จามจุรีโปรดักท์, 2009.

[12] Sokolova M., Lapalme G., Performance Measures in Classification of Human Communications.,Advances in Artificial Intelligence. AI 2007. Lecture Notes in Computer Science, vol 4509. Springer,Berlin, Heidelberg, 2007.

[13] George D. M., Andriana P., Machine Learning in Medical Applications, Springer-Verlag Berlin Heidelberg, 300-307, 2001.

[14] Frank E., Hall M., Holmes G., Kirkby R., Pfahringer B., Witten H. I., Trigg L., Weka-A Machine Learning Workbench for Data Mining, In: Maimon O., Rokach L. (eds) Data Mining and KnowledgeDiscovery Handbook. Springer, Boston, MA, 2010.

[15] Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I. H., Data mining in bioinformatics using Weka. Bioinformatics, 20(15), 2479–2481., 2004.

[16] Therneau T., Atkinson B., An Introduction to Recursive Partitioning Using the RPART Routines., https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf., 2018.

[17] Meyer D., Support Vector Machines—the Interface to libsvm in package e1071, https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf., 2018.

[18] Asri, H., Mousannif, H., Moatassime, H. A., Noel, T., Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science, 83, 1064–1069., 2016.

[19] Rau, H.-H., Hsu, C.-Y., Lin, Y.-A., Atique, S., Fuad, A., Wei, L.-M., Hsu, M.-H., Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. Computer Methods and Programs in Biomedicine, 125, 58–65., 2016.

[20] Cirkovic, B. R. A., Cvetkovic, A. M., Ninkovic, S. M., Filipovic, N. D,. Prediction models for estimation of survival rate and relapse for breast cancer patients., 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)., 2015.

[21] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR., Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. J Health Med Inform, 2013.

[22] Montazeri, M., Montazeri, M., Montazeri, M. ,Beigzadeh, A., Machine learning models in breast cancer survival prediction. Technology and Health Care, 24(1), 2016.

[23] S. Perveen., M. Shahbaz, K. Keshajee, A. Guergachi, A Systematic Machine Learning Based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression., Scientific reports, 2018.

[24] B. R. A. Cirkovic, Cvetkovic, A. M., Ninkovic, S. M., & Filipovic, N. D,. Prediction models for estimation of survival rate and relapse for breast cancer patients., 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)., 2015.

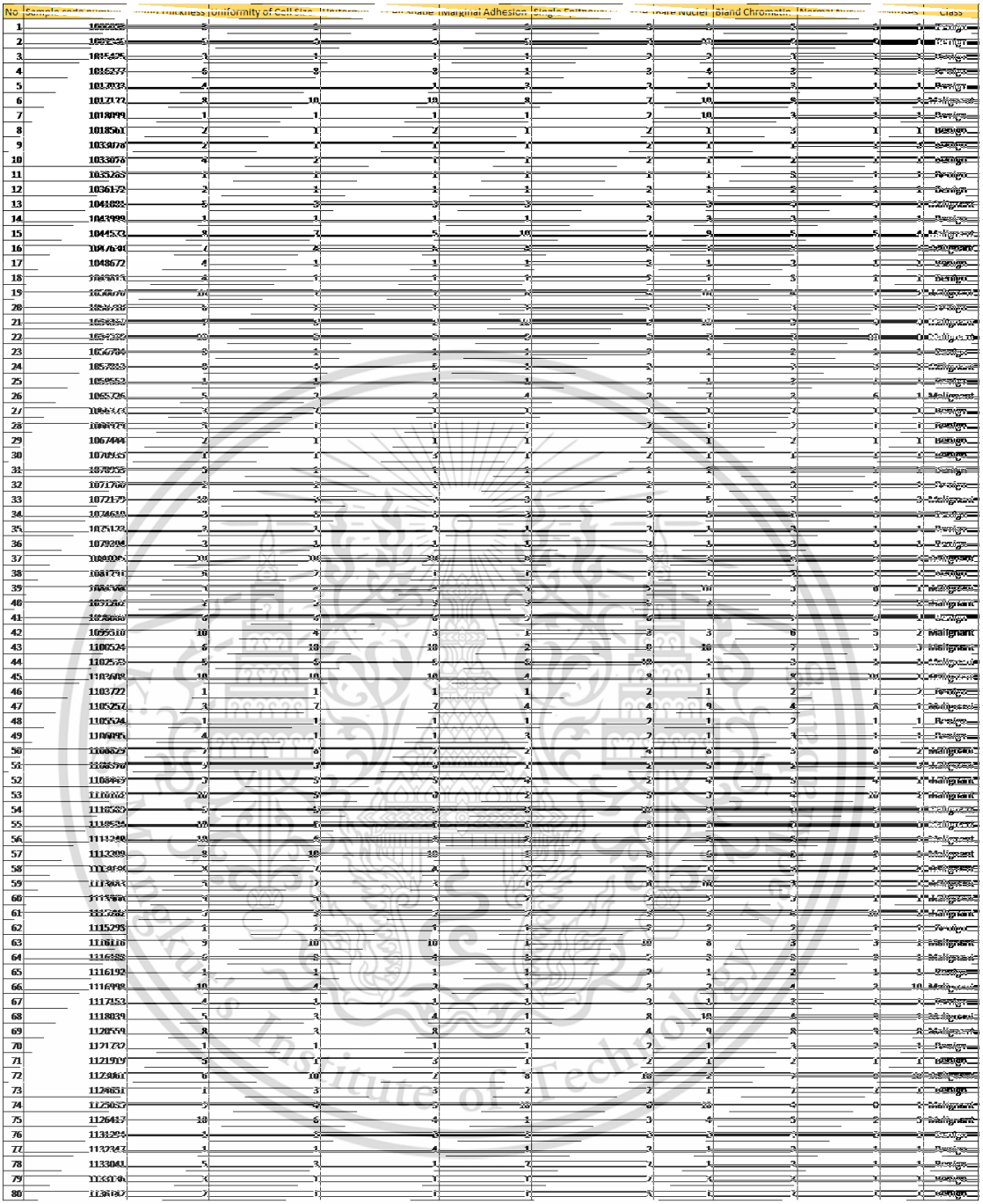
[25] Hamsagayathri, P., & Sampath, P. Priority based decision tree classifier for breast cancer detection. 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)., 2017.

[26] Chaurasia, V., Pal, S., & Tiwari, B. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, 119–126., 2018.

[27] Dr. WIlliam H. Wolberg, Olvi Mangasarian, University of Wisconsin Hospitals Madison, Wisconsin, USA, https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

Appendix A

Appendix A



No	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
81	1137156	2	2	2	1	1	1	7	1	1	Benign
82	1143978	4	1	1	2	2	1	2	1	1	Benign
83	1143978	5	2	1	1	2	1	3	1	1	Benign
84	1147044	3	1	1	1	2	2	7	1	1	Benign
85	1147699	3	5	7	8	8	9	7	10	7	Malignant
86	1147748	5	10	6	1	10	4	4	10	10	Malignant
87	1148278	3	3	6	4	5	8	4	4	1	Malignant
88	1148873	3	6	6	6	5	10	6	8	3	Malignant
89	1152331	4	1	1	1	2	1	3	1	1	Benign
90	1155546	2	1	1	2	3	1	2	1	1	Benign
91	1156272	1	1	1	1	2	1	3	1	1	Benign
92	1156948	3	1	1	2	2	1	1	1	1	Benign
93	1157734	4	1	1	1	2	1	3	1	1	Benign
94	1158247	1	1	1	1	2	1	2	1	1	Benign
95	1160476	2	1	1	1	2	1	3	1	1	Benign
96	1164066	1	1	1	1	2	1	3	1	1	Benign
97	1165297	2	1	1	2	2	1	1	1	1	Benign
98	1165790	5	1	1	1	2	1	3	1	1	Benign
99	1165926	9	6	9	2	10	6	2	9	10	Malignant
100	1166630	7	5	6	10	5	10	7	9	4	Malignant
101	1166654	10	3	5	1	10	5	3	10	2	Malignant
102	1167439	2	3	4	4	2	5	2	5	1	Malignant
103	1167471	4	1	2	1	2	1	3	1	1	Benign
104	1168359	8	2	3	1	6	3	7	1	1	Malignant
105	1168736	10	10	10	10	10	1	8	8	8	Malignant
106	1169049	7	3	4	4	3	3	3	2	7	Malignant
107	1170419	10	10	10	8	2	10	4	1	1	Malignant
108	1170420	1	6	8	10	8	10	5	7	1	Malignant
109	1171710	1	1	1	1	2	1	2	3	1	Benign
110	1171710	6	5	4	4	3	9	7	8	3	Malignant
111	1171795	1	3	1	2	2	2	5	3	2	Benign
112	1171845	8	6	4	3	5	9	3	1	1	Malignant
113	1172152	10	3	3	10	2	10	7	3	3	Malignant
114	1173216	10	10	10	3	10	[illegible]	[illegible]	[illegible]	[illegible]	Malignant
115	1173235	3	3	2	1	2	3	3	1	1	Benign
116	1173347	1	1	1	1	2	5	1	1	1	Benign
117	1173347	8	3	3	1	2	2	3	2	1	Benign
118	1173509	4	5	5	10	4	10	7	5	8	Malignant
119	1173514	1	1	1	1	4	3	1	1	1	Benign
120	1173681	3	2	1	1	2	2	3	1	1	Benign
121	1174057	1	1	2	2	2	1	3	1	1	Benign
122	1174057	4	2	1	1	2	2	3	1	1	Benign
123	1174131	10	10	10	2	10	10	5	3	3	Malignant
124	1174428	5	3	5	1	8	10	5	3	1	Malignant
125	1175937	5	4	6	7	9	7	8	10	1	Malignant
126	1176406	1	1	1	1	2	1	2	1	1	Benign
127	1176881	7	5	3	7	4	10	7	5	5	Malignant
128	1177027	3	1	1	1	2	1	3	1	1	Benign
129	1177399	8	3	5	4	5	10	1	6	2	Malignant
130	1177512	1	1	1	1	10	1	1	1	1	Benign
131	1178580	5	1	3	1	2	1	2	1	1	Benign
132	1179818	2	1	1	1	2	1	3	1	1	Benign
133	1180194	5	10	8	10	8	10	3	6	3	Malignant
134	1180523	3	1	1	1	2	1	2	2	1	Benign
135	1180831	3	1	1	1	3	1	2	1	1	Benign
136	1181356	5	1	1	1	2	2	3	3	1	Benign
137	1182404	4	1	1	1	2	1	2	1	1	Benign
138	1182410	3	1	1	1	2	1	1	1	1	Benign
139	1183240	4	1	2	1	2	1	2	1	1	Benign
140	1183246	1	1	1	1	1		2	1	1	Benign
141	1183516	3	1	1	1	2	1	1	1	1	Benign
142	1183911	2	1	1	1	2	1	1	1	1	Benign
143	1183983	9	5	5	4	4	5	4	3	3	Malignant
144	1184184	1	1	1	1	2	5	1	1	1	Benign
145	1184241	2	1	1	1	2	1	2	1	1	Benign
146	1184840	1	1	3	1	2		2	1	1	Benign
147	1185609	3	4	5	2	6	8	4	1	1	Malignant
148	1185610	1	1	1	1	3	2	2	1	1	Benign
149	1187457	3	1	1	3	8	1	5	8	1	Benign
150	1187805	8	8	7	4	10	10	7	8	7	Malignant
151	1188472	1	1	1	1	1	1	3	1	1	Benign
152	1189266	7	2	4	1	6	10	5	4	3	Malignant
153	1189286	10	10	8	6	4	5	8	10	1	Malignant
154	1190394	4	1	1	1	2	3	1	1	1	Benign
155	1190485	1	1	1	1	2	1	1	1	1	Benign
156	1192325	5	5	5	6	3	10	3	1	1	Malignant
157	1193091	1	2	2	1	2	1	2	1	1	Benign
158	1193210	2	1	1	1	2	1	3	1	1	Benign
159	1193683	1	1	2	1	3		1	1	1	Benign
160	1196295	9	9	10	3	6	10	7	10	6	Malignant
161	1196915	10	7	7	4	5	10	5	7	2	Malignant
162	1197080	4	1	1	1	2	1	3	2	1	Benign
163	1197270	3	1	1	1	2	1	3	1	1	Benign
164	1197440	1	1	1	2	1	3	1	1	7	Benign
165	1197510	5	1	1	1	2		3	1	1	Benign
166	1197979	4	1	1	1	2	2	3	2	1	Benign
167	1197993	5	6	7	8	8	10	3	10	3	Malignant
168	1198128	10	8	10	10	6	1	3	1	10	Malignant
169	1198641	3	1	1	1	2	1	3	1	1	Benign
170	1199219	1	1	1	2	1	1	1	1	1	Benign
171	1199731	3	1	1	1	2	1	1	1	1	Benign
172	1199983	1	1	1	1	2	1	3	1	1	Benign
173	1200772	1	1	1	1	2	1	2	1	1	Benign
174	1200847	6	10	10	10	8	10	10	10	7	Malignant
175	1200892	8	6	5	4	3	10	6	1	1	Malignant
176	1200952	5	8	7	7	10	10	5	7	1	Malignant
177	1201834	2	1	1	1	2	1	3	1	1	Benign
178	1201936	5	10	10	3	8	1	5	10	3	Malignant
179	1202125	4	1	1	1	2	1	3	1	1	Benign
180	1202812	5	3	3	3	6	10	3	1	1	Malignant

No	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
181	1202253	5	3	3	3	2	3	4	4	1	Malignant
182	1204002	4	3	4	1	2	3	2	3	1	Benign
183	1204898	6	1	1	1	2	1	2	1	1	Benign
184	1205138	5	8	8	2	5	10	7	8	1	Malignant
185	1205579	8	7	6	4	4	10	5	1	1	Malignant
186	1206089	2	1	1	1	1	1	1	1	1	Benign
187	1206841	1	1	1	1	2	1	2	1	1	Benign
188	1207986	10	3	8	10	8	10	7	10	1	Malignant
189	1207987	4	8	7	10	9	10	10	10	1	Malignant
190	1208301	1	2	1	1	2	1	3	1	1	Benign
191	1210979	10	10	10	8	8	10	7	10	1	Malignant
192	1211265	3	1	1	1	2	1	3	1	1	Benign
193	1211594	5	1	1	1	2	1	2	1	1	Benign
194	1211998	1	1	1	1	2	1	3	1	1	Benign
195	1212422	3	1	1	1	1	1	2	1	1	Benign
196	1212422	4	1	1	1	2	1	2	1	1	Benign
197	1213375	8	4	5	1	2	7	7	8	1	Malignant
198	1213383	5	1	1	1	2	1	3	1	1	Benign
199	1214092	1	1	1	1	2	1	1	1	1	Benign
200	1214301	3	2	2	1	2	1	2	1	1	Benign
201	1214556	3	1	1	1	2	10	3	1	1	Malignant
202	1214966	10	10	10	8	6	10	10	10	1	Malignant
203	1216694	1	1	1	1	2	1	2	1	1	Benign
204	1217051	5	1	1	1	2	1	2	1	1	Benign
205	1217264	1	1	1	1	2	1	2	1	1	Benign
206	1218105	5	10	10	10	6	10	7	10	1	Malignant
207	1218741	10	10	10	10	6	1	9	9	1	Malignant
208	1218860	1	1	1	1	1	1	3	1	1	Benign
209	1218860	1	1	1	1	1	1	3	1	1	Benign
210	1219406	5	1	1	1	1	1	3	1	1	Benign
211	1219525	8	10	10	10	8	10	8	10	1	Malignant
212	1219651	4	10	10	10	8	10	10	10	1	Malignant
213	1220330	1	1	1	1	2	1	2	1	1	Benign
214	1221863	10	10	10	10	10	10	10	10	1	Malignant
215	1222047	10	10	10	10	8	10	10	8	1	Malignant
216	1222936	8	7	8	7	5	10	7	10	1	Malignant
217	1223282	1	1	1	1	2	1	2	1	1	Benign
218	1223426	1	1	1	1	2	1	3	1	1	Benign
219	1223544	8	10	10	1	3	10	7	8	1	Malignant
220	1223567	5	1	1	1	2	1	3	1	1	Benign
221	1223769	2	1	1	1	2	1	3	1	1	Benign
222	1225799	10	6	4	3	3	10	3	4	2	Malignant
223	1226012	4	1	1	3	1	5	2	1	1	Malignant
224	1226612	7	5	6	3	3	8	7	4	1	Malignant
225	1227210	10	5	5	6	3	10	7	5	1	Malignant
226	1227244	1	1	1	1	2	1	3	1	1	Benign
227	1227481	10	5	7	4	4	10	8	2	1	Malignant
228	1227909	3	1	1	1	2	1	3	1	1	Benign
229	1228311	1	1	1	1	1	1	3	1	1	Benign
230	1228510	10	10	10	10	10	10	8	10	1	Malignant
231	1228698	2	3	2	1	2	1	2	1	1	Benign
232	1228807	3	1	1	1	2	1	3	1	1	Benign
233	1231706	4	2	3	1	2	1	3	1	1	Benign
234	1232225	10	4	5	5	5	10	4	3	1	Malignant
235	1236843	3	1	1	1	2	1	3	1	1	Benign
236	1241232	3	1	4	1	2	1	3	1	1	Benign
237	1241559	10	8	8	1	3	10	9	10	1	Malignant
238	1243256	4	2	3	1	2	1	2	1	1	Benign
239	1243674	5	10	10	3	5	8	10	7	1	Malignant
240	1243776	10	6	4	1	3	10	4	1	1	Malignant
241	1244158	4	1	2	1	2	1	2	1	1	Benign
242	1244270	5	1	1	1	2	1	3	1	1	Benign
243	1245718	2	1	1	1	2	1	3	1	1	Benign
244	1246562	2	2	2	2	2	1	2	2	1	Benign
245	1247533	1	1	1	1	2	1	2	1	1	Benign
246	1247975	6	1	1	1	2	1	2	1	1	Benign
247	1248627	8	10	10	8	5	10	7	8	1	Malignant
248	1248720	8	4	4	1	2	9	3	3	1	Malignant
249	1249406	4	1	1	1	2	1	2	1	1	Benign
250	1249550	5	1	1	1	2	1	3	1	1	Benign
251	1250525	3	2	2	1	2	1	1	1	1	Benign
252	1250644	10	4	4	10	2	10	7	5	1	Malignant
253	1251979	8	3	3	1	3	10	3	1	1	Benign
254	1252404	10	10	10	2	8	10	2	1	1	Malignant
255	1252627	8	10	10	2	10	10	5	3	1	Malignant
256	1253955	5	3	3	2	3	10	7	1	1	Malignant
257	1254526	2	1	1	1	2	1	3	1	1	Benign
258	1256272	3	1	1	1	2	1	3	1	1	Benign
259	1256836	1	1	1	1	2	1	2	1	1	Benign
260	1257648	5	7	7	1	3	8	3	2	1	Benign
261	1257815	10	3	6	10	5	10	5	10	1	Malignant
262	1257938	3	10	10	3	8	10	10	2	1	Malignant
263	1258049	8	6	7	1	3	10	7	1	1	Malignant
264	1258556	10	4	4	10	6	10	5	5	1	Malignant
265	1258577	5	3	5	1	8	10	5	1	1	Malignant
266	1258590	5	3	3	1	3	3	3	3	1	Benign
267	1258630	10	10	6	3	3	10	4	3	1	Malignant
268	1258741	3	3	5	3	3	10	7	1	1	Malignant
269	1258778	10	8	8	1	3	4	8	1	1	Malignant
270	1258778	1	1	1	1	2	1	3	1	1	Benign
271	1258778	8	4	7	1	3	10	3	3	1	Malignant
272	1258778	5	1	1	1	2	1	3	1	1	Benign
273	1258778	5	3	3	1	3	10	4	1	1	Malignant
274	1258778	7	2	4	1	3	4	5	3	1	Malignant
275	1258778	3	2	2	1	2	2	2	1	1	Benign
276	1258778	3	2	2	2	2	1	2	2	1	Benign
277	1258778	3	1	1	1	2	1	2	1	1	Benign
278	1258778	1	1	1	1	2	1	3	1	1	Benign
279	1258778	1	1	1	1	1	1	2	1	1	Benign
280	1258778	10	3	4	3	3	1	4	3	1	Malignant

No	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
281	488827	2	1	1	1	2	1	2	1	1	Benign
282	489726	5	3	3	3	2	3	3	1	1	Benign
283	489931	1	4	1	10	4	10	7	2	1	Malignant
284	492268	10	4	6	1	2	10	7	2	1	Malignant
285	508234	7	4	5	10	7	10	3	8	1	Malignant
286	527363	10	10	10	10	10	10	10	10	1	Malignant
287	529329	10	10	10	10	10	10	8	8	1	Malignant
288	533548	5	1	1	1	3	1	2	1	1	Benign
289	545328	8	4	5	1	4	5	5	10	1	Malignant
290	555977	5	6	6	3	6	10	6	10	1	Malignant
291	560680	3	1	1	1	2	1	3	1	1	Benign
292	561477	1	1	1	1	2	1	2	1	1	Benign
293	563649	8	8	8	1	3	5	8	10	1	Malignant
294	601265	10	4	4	6	2	10	3	2	1	Malignant
295	606140	1	1	1	1	1	1	3	1	1	Benign
296	606722	5	3	4	3	4	10	6	2	1	Malignant
297	616240	5	3	4	3	4	5	5	1	1	Benign
298	61634	3	1	1	1	2	1	3	1	1	Benign
299	625201	8	4	6	3	4	10	6	10	1	Malignant
300	63375	9	5	7	5	5	10	7	10	1	Malignant
301	635844	8	4	10	3	4	10	7	10	1	Malignant
302	636130	5	2	3	1	2	3	5	1	1	Benign
303	640744	10	10	10	7	9	10	7	10	1	Malignant
304	644171	1	1	1	1	3	1	3	1	1	Benign
305	653777	8	8	9	8	5	10	7	1	1	Malignant
306	659642	10	6	4	4	4	10	7	10	1	Malignant
307	666090	1	1	1	1	1	1	1	1	1	Benign
308	666942	1	1	1	1	2	1	3	1	1	Benign
309	667204	7	8	7	6	4	8	8	9	1	Malignant
310	673637	3	1	1	1	2	1	3	1	1	Benign
311	684955	2	1	1	1	3	1	2	1	1	Benign
312	688033	2	1	1	1	2	1	1	1	1	Benign
313	691628	8	6	6	10	10	10	6	5	2	Malignant
314	693702	1	1	1	1	2	1	1	1	1	Benign
315	704097	1	1	1	1	1	1	2	1	1	Benign
316	704168	4	6	5	7	6	10	7	7	2	Malignant
317	706426	5	5	5	2	5	10	8	10	1	Malignant
318	709287	8	8	7	3	4	8	3	8	1	Malignant
319	718641	1	1	1	1	2	1	1	1	1	Benign
320	721482	4	4	4	4	6	4	7	7	2	Malignant
321	730881	7	6	5	10	7	10	5	3	1	Malignant
322	733639	3	1	1	1	2	1	3	1	1	Benign
323	733640	3	1	1	1	2	3	3	1	1	Benign
324	733639	8	8	9	10	3	10	7	5	1	Malignant
325	740492	1	1	1	1	2	1	2	1	1	Benign
326	743348	3	2	2	1	2	1	2	1	1	Benign
327	752904	10	1	1	1	2	10	5	2	1	Malignant
328	756136	1	1	1	1	2	1	2	1	1	Benign
329	760001	8	10	3	2	10	4	3	10	1	Malignant
330	760239	10	4	6	4	5	10	7	5	1	Malignant
331	76389	10	4	7	2	2	8	6	1	1	Malignant
332	764974	3	1	1	1	2	1	3	1	1	Benign
333	770066	5	2	2	2	2	1	2	1	1	Benign
334	785208	5	5	6	10	5	10	5	10	1	Malignant
335	785615	8	7	8	7	3	10	9	1	1	Malignant
336	797744	1	1	1	1	2	1	3	1	1	Benign
337	797327	5	5	5	8	8	10	5	7	1	Malignant
338	798429	1	1	1	1	2	1	3	1	1	Benign
339	704097	1	1	1	1	1	1	2	1	1	Benign
340	806423	8	5	5	5	2	10	6	3	1	Malignant
341	809912	10	3	3	1	2	10	7	2	1	Malignant
342	810104	1	1	1	1	2	1	3	1	1	Benign
343	814265	2	1	1	1	2	1	3	1	1	Benign
344	814366	3	1	2	1	2	1	3	1	1	Benign
345	822829	8	6	4	6	4	10	6	3	1	Malignant
346	826923	1	1	1	1	2	1	3	1	1	Benign
347	830690	5	3	3	3	2	3	4	1	1	Benign
348	831268	1	1	1	1	2	1	3	1	1	Benign
349	832226	3	4	4	10	5	1	3	3	1	Malignant
350	832567	4	2	3	5	3	8	3	4	1	Malignant
351	836433	3	1	1	3	2	1	1	1	1	Benign
352	837082	2	4	4	1	2	1	2	1	1	Benign
353	846832	3	4	5	5	7	3	4	7	1	Benign
354	850831	2	7	10	10	7	10	4	5	1	Malignant
355	855524	1	1	1	1	2	1	2	1	1	Benign
356	857774	4	6	5	6	7	7	8	2	1	Malignant
357	859164	5	3	3	1	3	3	6	1	1	Benign
358	859350	8	10	10	7	10	10	7	3	1	Malignant
359	866325	8	10	7	3	8	10	9	10	1	Malignant
360	873843	10	7	7	4	5	7	5	6	1	Malignant
361	877291	6	10	10	10	10	10	8	10	1	Malignant
362	877943	3	10	7	10	7	10	7	5	1	Malignant
363	888169	3	2	2	1	4	1	3	2	1	Benign
364	888523	4	4	4	2	3	3	5	3	1	Benign
365	896404	2	1	2	1	2	1	2	1	1	Benign
366	897471	4	8	8	5	4	5	5	10	1	Malignant
367	897604	6	10	10	10	8	10	7	10	1	Malignant
368	160296	5	8	8	10	5	10	8	10	3	Malignant
369	342245	1	1	3	1	2	1	1	1	1	Benign
370	428903	1	1	1	1	1	1	2	1	1	Benign
371	492561	4	3	2	1	3	1	2	1	1	Benign
372	493452	1	1	3	1	2	1	1	1	1	Benign
373	493452	4	1	2	1	2	1	2	1	1	Benign
374	521441	5	1	1	2	2	1	2	1	1	Benign
375	560680	3	1	1	1	2	1	3	1	1	Benign
376	636437	1	1	1	1	2	1	3	1	1	Benign
377	640744	10	10	10	7	9	10	7	10	1	Malignant
378	646904	1	1	1	1	2	1	3	1	1	Benign
379	657753	3	1	1	4	3	1	3	2	1	Benign
380	685977	5	8	8	10	5	10	7	10	1	Malignant

No	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
381	869222	2	2	2	2	2	2	2	2	1	Benign
382	866325	10	7	5	3	4	10	7	1	2	Malignant
383	1003532	2	1	1	1	2	1	2	1	1	Benign
384	1003852	2	1	1	1	2	1	1	1	1	Benign
385	1026122	2	1	1	1	2	1	1	1	1	Benign
386	1027029	1	1	1	1	1	1	1	1	1	Benign
387	1033733	7	8	8	5	7	10	9	7	1	Malignant
388	1035670	3	1	1	1	2	1	2	1	1	Benign
389	1036172	2	1	1	1	2	1	2	1	1	Benign
390	1037227	3	1	1	1	3	2	1	1	1	Benign
391	1038611	1	1	1	1	2	1	2	1	1	Benign
392	1041334	10	8	7	5	7	10	7	1	1	Malignant
393	1043068	3	1	1	1	2	2	2	1	1	Benign
394	1044572	1	1	1	1	1	1	1	1	1	Benign
395	1047630	3	2	3	1	2	1	3	1	1	Benign
396	1048672	1	1	1	1	1	1	1	1	1	Benign
397	1049815	4	1	1	1	2	1	3	1	1	Benign
398	1050670	10	7	7	10	7	10	4	1	2	Malignant
399	1050718	3	1	1	1	2	1	3	1	1	Benign
400	1054590	3	1	1	1	2	1	2	1	1	Benign
401	1054593	10	10	10	10	6	10	10	10	2	Malignant
402	1056784	2	1	1	1	2	1	1	1	1	Benign
403	1057013	8	4	5	1	2	10	4	1	2	Malignant
404	1059552	1	1	1	1	2	1	1	1	1	Benign
405	1065073	1	1	1	1	2	1	1	1	1	Benign
406	1066373	1	2	1	1	2	1	1	1	1	Benign
407	1066979	5	1	1	1	2	1	2	1	1	Benign
408	1067444	2	1	1	1	2	1	1	1	1	Benign
409	1070935	1	1	1	1	2	1	1	1	1	Benign
410	1070935	3	1	1	1	2	1	2	1	1	Benign
411	1071760	2	1	1	1	2	1	2	1	1	Benign
412	1072179	10	7	7	3	8	5	7	1	1	Malignant
413	1074465	2	1	1	1	2	1	3	1	1	Benign
414	1077001	1	1	1	1	2	1	1	1	1	Benign
415	1077258	4	1	1	1	2	1	2	1	1	Benign
416	1077475	2	1	1	1	2	1	1	1	1	Benign
417	1078612	5	3	3	2	3	1	1	1	1	Benign
418	1079304	2	1	1	1	2	1	1	1	1	Benign
419	1080185	10	10	10	8	6	10	10	10	1	Malignant
420	1081791	5	1	1	1	2	1	3	1	1	Benign
421	1084584	5	1	1	1	2	1	2	1	1	Benign
422	1091262	5	2	2	2	2	1	2	1	1	Benign
423	1096352	6	10	10	10	8	10	7	10	3	Malignant
424	1100524	6	2	1	1	2	1	1	1	1	Benign
425	1102573	5	1	1	1	2	1	3	1	1	Benign
426	1103608	10	10	10	10	8	10	7	10	1	Malignant
427	1103722	2	1	1	1	2	1	2	1	1	Benign
428	1105257	3	1	1	1	2	1	1	1	1	Benign
429	1105524	4	1	1	1	2	1	1	1	1	Benign
430	1106095	4	1	1	1	2	1	2	1	1	Benign
431	1106829	1	1	1	1	2	1	2	1	1	Benign
432	1108370	9	5	8	1	2	3	2	1	5	Malignant
433	1108449	5	3	3	3	2	3	4	1	1	Benign
434	1110102	10	3	6	2	3	5	4	10	2	Malignant
435	1110503	5	3	4	4	4	5	6	1	1	Benign
436	1110524	10	5	5	3	6	7	7	10	1	Malignant
437	1112209	10	10	10	8	5	10	7	10	1	Malignant
438	1113038	8	4	5	1	2	5	5	10	4	Malignant
439	1113483	5	1	2	3	2	4	3	10	1	Malignant
440	1113906	3	1	1	1	2	1	1	1	1	Benign
441	1115282	9	5	7	7	9	10	10	1	1	Malignant
442	1116116	5	7	7	1	3	8	3	8	2	Malignant
443	1116132	5	1	2	1	2	1	2	1	1	Benign
444	1116192	2	1	1	1	2	1	3	1	1	Benign
445	1116998	5	3	3	3	2	3	4	1	1	Benign
446	1117152	3	1	1	1	2	1	1	1	1	Benign
447	1118039	5	3	4	1	8	10	4	1	2	Malignant
448	1120559	3	1	1	1	2	1	3	1	1	Benign
449	1121732	1	1	1	1	2	1	3	1	1	Benign
450	1121919	5	1	3	1	2	1	1	1	1	Benign
451	1122965	8	10	10	8	7	10	7	1	2	Malignant
452	1123061	6	1	1	1	2	1	3	1	1	Benign
453	1124651	3	1	1	3	2	1	1	1	1	Benign
454	1125035	9	8	6	4	6	10	6	1	3	Malignant
455	1126417	2	1	1	1	2	1	2	1	1	Benign
456	1131294	1	1	2	1	2	1	3	1	1	Benign
457	1132347	1	1	1	1	2	1	1	1	1	Benign
458	1133041	5	3	3	3	2	3	4	1	1	Benign
459	1133136	10	10	10	10	6	10	10	10	10	Malignant
460	1135232	5	7	7	1	5	8	8	2	1	Malignant
461	1136142	2	1	1	1	2	1	1	1	1	Benign
462	1137156	2	1	1	1	2	1	2	1	1	Benign
463	1137974	5	10	10	3	7	3	8	10	2	Malignant
464	1138404	4	1	1	1	2	1	2	1	1	Benign
465	1138467	1	1	1	1	2	1	1	1	1	Benign
466	1138666	10	8	7	5	6	4	7	10	3	Malignant
467	1139536	2	1	1	1	2	1	2	1	1	Benign
468	1139948	10	6	6	7	4	10	9	1	2	Malignant
469	1140132	5	3	3	3	2	3	4	1	1	Benign
470	1141129	1	1	1	1	2	1	3	1	1	Benign
471	1141259	8	10	10	10	6	10	10	10	4	Malignant
472	1142706	10	10	10	10	5	10	10	10	2	Malignant
473	1143978	4	1	1	1	2	1	2	1	1	Benign
474	1147044	3	1	1	1	2	1	3	1	1	Benign
475	1147699	3	1	1	1	2	1	3	1	1	Benign
476	1147748	5	10	6	10	4	10	7	1	2	Malignant
477	1148278	3	1	1	1	2	1	3	2	1	Benign
478	1148873	2	1	1	1	2	1	3	1	1	Benign
479	1152331	4	1	1	1	2	1	3	1	1	Benign
480	1153089	4	1	1	3	2	1	3	1	1	Benign

No	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
491	1293439	3	2	2	1	2	1	1	1	1	Benign
492	1294562	5	2	2	2	2	2	2	1	1	Benign
493	1295186	9	10	10	10	10	5	10	10	10	Malignant
494	1298416	7	7	7	5	5	10	6	10	10	Malignant
495	1299596	5	1	2	1	2	1	1	1	1	Benign
496	1299994	5	3	3	2	3	1	1	2	1	Benign
497	1301264	3	1	1	1	2	1	1	1	1	Benign
498	1302347	5	1	1	1	2	1	1	1	1	Benign
499	1304595	3	1	1	1	2	1	2	1	1	Benign
500	1306339	4	1	1	1	2	1	2	1	1	Benign
501	1306779	5	2	2	2	2	1	2	1	1	Benign
502	1307556	3	1	1	1	2	1	1	1	1	Benign
503	1307604	3	1	1	1	2	1	2	1	1	Benign
504	1307663	3	1	1	1	2	1	3	1	1	Benign
505	1307675	3	1	1	1	2	1	1	1	1	Benign
506	1307681	5	1	1	1	2	1	1	1	1	Benign
507	1307704	10	10	10	10	4	10	7	10	1	Malignant
508	1307718	3	1	1	1	2	1	1	1	1	Benign
509	1307992	3	1	1	1	2	1	2	1	1	Benign
510	1308037	2	1	1	1	2	1	1	1	1	Benign
511	1308589	4	2	1	1	2	2	2	2	1	Benign
512	1308951	5	2	2	2	2	1	2	2	1	Benign
513	1308964	5	2	2	2	2	2	2	2	1	Benign
514	1309420	3	1	1	1	2	1	1	1	1	Benign
515	1309539	5	8	8	1	2	10	9	10	1	Malignant
516	1312235	8	10	8	1	2	10	9	10	1	Malignant
517	1312611	3	1	1	1	2	1	1	1	1	Benign
518	1312675	3	1	1	1	2	1	1	1	1	Benign
519	1313658	3	1	2	1	2	1	3	1	1	Benign
520	1313982	4	1	1	3	2	10	3	1	1	Malignant
521	1314427	5	1	2	1	2	1	1	1	1	Benign
522	1315506	4	2	2	2	2	2	2	2	1	Benign
523	1320141	10	4	5	3	3	10	3	5	2	Malignant
524	1320672	5	5	6	10	6	10	7	7	10	Malignant
525	1320843	3	1	1	1	2	1	2	1	1	Benign
526	1320959	3	1	1	2	2	1	1	1	1	Benign
527	1321264	5	1	1	1	2	1	1	1	1	Benign
528	1321321	4	1	1	1	2	1	2	1	1	Benign
529	1321348	10	6	5	7	7	10	5	5	3	Malignant
530	1321931	4	1	1	1	1	1	2	1	1	Benign
531	1321942	7	4	4	3	4	10	6	10	2	Malignant
532	1321956	5	2	2	2	2	2	2	2	1	Benign
533	1322359	2	1	1	1	2	1	2	1	1	Benign
534	1322477	5	1	1	1	2	1	2	1	1	Benign
535	1322548	3	1	1	1	2	1	1	1	1	Benign
536	1322818	3	1	2	2	2	1	3	1	1	Benign
537	1322829	5	3	3	1	3	1	3	1	1	Benign
538	1322865	5	1	1	1	2	1	1	1	1	Benign
539	1323256	3	1	1	1	2	1	1	1	1	Benign
540	1323309	4	1	1	1	2	1	2	1	1	Benign
541	1323577	5	2	2	2	2	2	2	1	1	Benign
542	1324348	5	2	2	2	2	2	2	2	1	Benign
543	1324572	5	3	3	2	3	2	3	1	1	Benign
544	1324681	3	1	1	1	2	1	2	1	1	Benign
545	1326337	2	1	2	2	2	1	2	1	1	Benign
546	1326348	5	1	1	1	2	1	1	1	1	Benign
547	1326521	5	10	10	10	4	10	7	10	1	Malignant
548	1326892	3	1	1	1	1	1	1	1	1	Benign
549	1326895	3	1	1	1	1	1	1	1	1	Benign
550	1326928	7	8	8	7	6	8	7	8	2	Malignant
551	1327091	5	1	1	1	2	1	2	1	1	Benign
552	1327139	3	1	1	1	2	1	3	1	1	Benign
553	1327141	3	2	2	2	2	1	3	2	1	Benign
554	1327157	3	2	2	2	2	2	2	2	1	Benign
555	1327270	3	2	2	2	2	2	2	2	1	Benign
556	1327327	3	2	3	2	3	2	3	2	1	Benign
557	1327382	5	3	3	3	3	1	3	1	1	Benign
558	1327395	5	1	1	3	2	1	1	1	1	Benign
559	1327651	4	1	1	1	2	1	2	1	1	Benign
560	1327780	3	1	1	1	2	1	2	1	1	Benign
561	1327836	4	1	1	1	2	1	3	1	1	Benign
562	1328182	3	1	1	1	2	1	3	1	1	Benign
563	1328222	4	2	2	2	2	2	2	2	1	Benign
564	1328755	3	1	1	1	2	1	3	1	1	Benign
565	1328948	4	2	2	2	2	1	2	2	1	Benign
566	1328982	5	7	10	10	5	10	10	10	1	Malignant
567	1328997	3	2	2	2	2	2	2	2	1	Benign
568	1329274	4	1	1	1	2	3	3	1	1	Benign
569	1329550	8	4	4	1	6	10	2	5	2	Malignant
570	1329641	10	10	10	10	10	10	10	10	3	Malignant
571	1329710	8	10	4	4	8	10	8	10	1	Malignant
572	1329896	7	6	10	5	5	10	9	10	2	Malignant
573	1330439	5	1	1	1	2	1	2	1	1	Benign
574	1330637	3	2	2	2	2	2	2	2	1	Benign
575	1331071	10	3	3	3	4	10	7	9	2	Malignant
576	1331387	5	2	2	2	2	2	2	2	1	Benign
577	1331709	5	3	3	3	3	3	3	3	1	Benign
578	1333063	1	1	1	1	2	1	3	1	1	Benign
579	1333220	1	1	1	1	2	1	2	1	1	Benign
580	1333349	3	1	1	1	2	1	3	1	1	Benign
581	1333495	3	1	1	1	2	1	3	1	1	Benign
582	1333849	3	7	10	8	7	10	7	1	1	Malignant
583	1334071	6	10	10	10	8	10	7	10	1	Malignant
584	1334223	3	3	3	3	3	3	3	3	1	Benign
585	1217717	5	1	1	1	2	1	1	1	1	Benign
586	1329439	3	1	1	1	2	2	3	1	1	Benign
587	1254539	8	10	10	10	6	10	10	10	1	Malignant
588	1261751	5	1	1	1	2	1	3	1	1	Benign
589	1268275	9	8	8	9	6	3	4	1	1	Malignant
590	1272166	5	1	1	1	2	1	1	1	1	Benign

No	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
591	1294352	4	10	6	6	4	10	10	5	1	Malignant
592	1295529	2	5	7	6	4	10	7	2	2	Malignant
593	1298484	10	3	4	5	3	10	4	1	1	Malignant
594	1311875	5	1	2	1	2	1	1	1	1	Benign
595	1315506	4	8	8	3	4	10	7	1	1	Malignant
596	1320141	5	1	1	1	2	1	2	1	1	Benign
597	1325309	4	1	2	1	2	1	2	1	1	Benign
598	1333063	3	1	3	1	2	1	3	1	1	Benign
599	1333495	3	1	2	2	2	1	2	1	1	Benign
600	1334659	5	2	4	2	2	2	4	1	1	Benign
601	1336798	5	2	3	1	2	2	2	1	1	Benign
602	1344449	3	1	1	1	2	1	1	1	1	Benign
603	1350568	4	1	1	1	2	1	2	1	1	Benign
604	1350774	5	4	6	8	4	1	8	10	1	Malignant
605	1352848	5	3	3	1	2	10	9	1	1	Malignant
606	1352904	10	7	10	7	7	10	7	1	1	Malignant
607	1354840	4	1	1	2	2	1	2	1	1	Benign
608	1354840	1	1	1	1	2	1	1	1	1	Benign
609	1355260	1	10	10	10	10	10	10	10	2	Malignant
610	1365075	5	1	1	1	2	1	1	1	1	Benign
611	1365328	10	5	5	10	6	10	5	2	1	Malignant
612	1368267	5	10	10	10	5	5	5	10	10	Malignant
613	1368273	1	1	1	1	2	1	1	1	1	Benign
614	1368882	2	1	1	1	2	1	1	1	1	Benign
615	1369821	2	1	1	1	2	1	2	1	1	Benign
616	1371026	4	1	2	1	2	1	2	1	1	Benign
617	1371041	3	1	1	1	2	1	2	1	1	Benign
618	1371920	1	1	1	1	1	1	1	1	1	Benign
619	1371926	4	1	1	1	2	1	2	1	1	Benign
620	1372568	3	1	1	1	2	1	2	1	1	Benign
621	1372818	5	1	1	1	2	1	2	1	1	Benign
622	1373836	8	5	5	5	5	2	6	1	1	Benign
623	1373943	7	2	2	2	2	2	2	2	1	Benign
624	1376178	4	1	1	1	2	1	1	1	1	Benign
625	1378275	5	1	1	1	2	1	3	1	1	Benign
626	1385610	2	1	2	1	2	1	2	1	1	Benign
627	1385942	8	8	8	1	3	4	3	7	1	Malignant
628	1386526	3	1	1	1	2	1	3	1	1	Benign
629	1390133	7	1	1	1	2	1	1	1	1	Benign
630	1390836	4	1	1	1	2	1	2	1	1	Benign
631	1391037	6	1	1	1	2	1	2	1	1	Benign
632	1391869	5	2	2	2	3	1	2	1	1	Benign
633	1393549	2	2	2	2	2	1	2	2	1	Benign
634	1393949	6	5	5	8	6	5	7	10	1	Malignant
635	1394235	2	1	1	1	2	1	1	1	1	Benign
636	1260659	3	1	1	1	2	1	1	1	1	Benign
637	1268952	10	10	7	8	7	1	10	10	3	Malignant
638	1275807	8	6	5	4	5	10	6	1	1	Malignant
639	1277792	4	1	1	1	2	1	1	1	1	Benign
640	1277792	5	1	1	3	2	1	1	1	1	Benign
641	1277792	4	1	1	1	2	1	2	1	1	Benign
642	1285531	3	1	1	1	2	1	2	1	1	Benign
643	1287775	5	3	3	3	2	3	4	4	1	Malignant
644	1287971	3	1	1	1	2	1	1	1	1	Benign
645	1292597	3	1	1	1	2	1	3	1	1	Benign
646	1308428	2	1	1	1	2	1	2	1	1	Benign
647	1311033	5	1	1	1	2	1	3	1	1	Benign
648	1311108	1	1	1	1	2	1	1	1	1	Benign
649	1315807	5	10	10	10	8	10	10	10	2	Malignant
650	1318671	1	1	1	1	2	1	2	1	1	Benign
651	1319609	3	1	1	1	2	3	1	1	1	Benign
652	1323477	1	1	1	1	2	1	1	1	1	Benign
653	1324572	5	1	1	1	2	1	2	1	1	Benign
654	1324681	4	1	1	1	2	1	2	1	1	Benign
655	1325159	2	1	1	1	2	1	2	1	1	Benign
656	1326249	3	1	1	1	2	1	2	1	1	Benign
657	1330361	5	1	1	1	2	1	2	1	1	Benign
658	1333877	5	4	5	1	8	1	5	6	1	Benign
659	1334071	4	10	10	1	6	10	5	2	2	Malignant
660	1334659	5	1	1	1	2	1	3	1	1	Benign
661	1336798	1	1	1	1	2	1	2	1	1	Benign
662	1339781	4	1	1	1	2	1	3	1	1	Benign
663	1342945	3	1	1	1	2	1	2	1	1	Benign
664	1342932	2	1	1	1	2	1	3	1	1	Benign
665	1344121	5	10	10	10	4	10	7	2	3	Malignant
666	1346478	5	1	1	1	2	1	2	1	1	Benign
667	1346482	5	1	1	1	2	1	3	1	1	Benign
668	1346832	3	1	1	1	2	1	3	1	1	Benign
669	1346852	1	1	1	1	2	1	3	1	1	Benign
670	1347972	5	10	10	5	5	10	8	1	2	Malignant
671	1347973	5	10	8	5	5	10	7	1	1	Malignant
672	1350424	3	1	1	1	2	1	2	1	1	Benign
673	1351346	2	1	1	1	2	1	2	1	1	Benign
674	1351609	3	1	1	1	2	1	2	1	1	Benign
675	1352529	3	1	1	1	2	1	2	1	1	Benign
676	1354840	4	1	1	1	2	1	2	1	1	Benign
677	1355328	3	1	1	1	2	1	2	1	1	Benign
678	1368267	5	1	1	1	2	1	1	1	1	Benign
679	1368273	1	1	1	1	2	1	1	1	1	Benign
680	1368882	2	1	1	1	2	1	1	1	1	Benign
681	1369821	10	10	10	10	5	10	10	10	7	Malignant
682	1371026	5	10	10	10	4	10	5	5	3	Malignant
683	1371920	1	1	1	1	2	1	3	1	1	Benign
684	466906	1	1	1	1	2	1	1	1	1	Benign
685	468973	1	1	1	1	2	1	2	1	1	Benign
686	524658	2	1	1	1	2	1	2	1	1	Benign
687	536708	1	1	1	1	2	1	1	1	1	Benign
688	566346	3	1	1	1	2	1	2	1	1	Benign
689	603148	4	1	1	1	2	1	1	1	1	Benign
690	654546	1	1	1	1	2	1	1	1	1	Benign
691	654546	1	1	1	1	2	1	1	1	1	Benign
692	695091	5	10	10	5	4	5	4	4	1	Malignant
693	714039	3	1	1	1	2	1	3	1	1	Benign
694	763235	3	1	1	1	2	1	3	1	1	Benign
695	776715	3	1	1	1	3	2	1	1	1	Benign
696	841769	2	1	1	1	2	1	1	1	1	Benign
697	888820	5	10	10	3	7	3	8	10	2	Malignant
698	897471	4	8	6	4	3	4	10	6	1	Malignant
699	897471	4	8	8	5	4	5	10	4	1	Malignant

Author Biography

Name	Miss Jiraphorn Charoenpong
Date of Birth	17 September 1994
Address	99/212, Kanjanapisek Road, Prawet, Prawet District Bangkok, 10250
Education	2017 Bachelor of Science in Applied Mathematics GPA 2.86 King Mongkut's Institute of Technology Ladkrabang
	2019 Master of Science in Applied Mathematics GPA 3.32 King Mongkut's Institute of Technology Ladkrabang
Scholarship(s)	Assistant Researcher by Centre of Excellence in Mathematics
	Assistant Researcher by Thailand Center of Excellence in Physics

Academic Publication(s)

1. Jiraporn C., Busayamas P., Somkid A., Wannapong T. and Narin N. "A Comparison of Machine Learning Algorithms and their Applications", International Journal of Simulation Systems, Science and Technology, Volume 20, Number 4, August 2019. DOI 10.5013/IJSSST.a.20.04.08