EMITTER International Journal of Engineering Technology
Vol. 5, No. 1, June 2017
ISSN: 2443-1168

Data Mining Approach for Breast Cancer Patient Recovery

Tresna Maulana Fahrudin, Iwan Syarif, Ali Ridho Barakbah

Department of Information and Computer Engineering
Graduate Program of Engineering Technology
Politeknik Elektronika Negeri Surabaya
Jl. Raya ITS Sukolilo Surabaya 60111, Indonesia
Telp: 6231 5947280 Fax : 6231 5946114
E-mail: tresnamf@pasca.student.pens.ac.id, {iwanarif, ridho}@pens.ac.id

Abstract

Breast cancer is the second highest cancer type which attacked Indonesian women. There are several factors known related to encourage an increased risk of breast cancer, but especially in Indonesia that factors often depends on the treatment routinely. This research examines the determinant factors of breast cancer and measures the breast cancer patient data to build the useful classification model using data mining approach.The dataset was originally taken from one of Oncology Hospital in East Java, Indonesia, which consists of 1097 samples, 21 attributes and 2 classes. We used three different feature selection algorithms which are Information Gain, Fisher's Discriminant Ratio and Chi-square to select the best attributes that have great contribution to the data. We applied Hierarchical K-means Clustering to remove attributes which have lowest contribution. Our experiment showed that only 14 of 21 original attributes have the highest contribution factor of the breast cancer data. The clustering algorithmdecreased the error ratio from 44.48% (using 21 original attributes) to 18.32% (using 14 most important attributes).We also applied the classification algorithm to build the classification model and measure the precision of breast cancer patient data. The comparison of classification algorithms between Naïve Bayes and Decision Tree were both given precision reach 92.76% and 92.99% respectively by leave-one-out cross validation. The information based on our data research, the breast cancer patient in Indonesia especially in East Java must be improved by the treatment routinely in the hospital to get early recover of breast cancer which it is related with adherence of patient.

Keywords: Data Mining, Breast Cancer, Feature Selection, Clustering, Classification.

1. INTRODUCTION

International Agency for Research on Cancer (IARC) through GLOBOCAN 2012 Projects reported that the diagnosis of cancer globally in 2012 discovered a type of cancer that contributed to the second highest mortality rate was breast cancer with percentage of 11.9% or 1.7 million women. If it was compared to the previous case of breast cancer in 2008 there was 6.3 million women diagnosed, there was 20% increase in new case and 14% of the total mortality of breast cancer patients. Breast cancer is also the most common cause of death among women with cancer (522.000 deaths in 2012) and type of cancer most attacked women in 140 of 184 countries in the world [11].

According to the demographic census, Indonesian women have longer life expectancy than ten years ago. Long life expectancy that is meaning a higher possibility leads to chronic diseases. The statistic from Ministry of Health [19] reported that the ranking of chronic disease morbidity and mortality was increased. Cancer ranks asthe fifth in the morbidity and the third in the mortality. Breast cancer ranks as the second after cervical cancer in women [8][22] with relative frequency 18% to the pathology and limited population based cancer registry.

According to the National Breast and Ovarian Cancer Centre (NBOCC), the several factors are known to relate to encourage an increased risk of breast cancer. It factors are related with an increased risk of breast cancer of woman generally whoare more attacked breast cancer than they without the disease. Several markers for suspected factors that influence risk is sex, age, affluence, family history, breast conditions, endogenous oestrogens, hormonal, personal and life styles [20]. There are still many factors that influence patients were attacked breast cancer with approach and research in different cases.

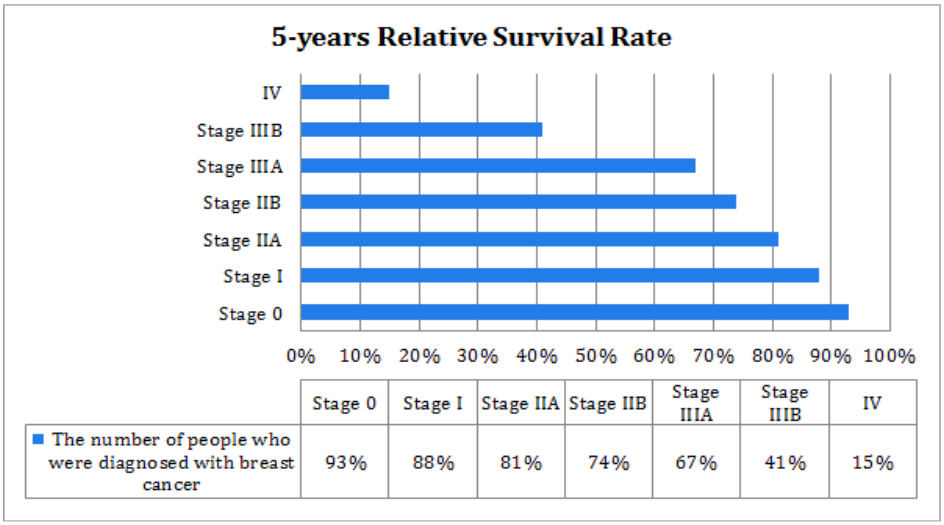

Figure 1. 5-years survival rate was reported by American Cancer Society,The numbers come from the National Cancer Data Base, and are based on people who were diagnosed with breast cancer in 2001 and 2002.

Survival rate is often used to indicate the result of breast cancer treatment. According to the American Cancer Society, 5-years survival rate (the percentage of patients who life at least 5 years after being diagnosed with breast cancer) that is shown in Figure 1 that stage 0 reached 93%, stage I reached 88%, stage IIA reached 81%, stage IIB reached 74%, stage IIIA reached 67%, stage IIIB reached 41%, and stage IV reached 15% [3]. In the United States of America, the mortality rate gradually decreased that has been related with breast cancer treatment progress. Survival rate in women has increased to 13% since the mid-1970s [15]. The different case, the mortality is still increasing in Indonesia and other developing countries, not only the late history of diseases, but also patient's non-adherence to treatment routinely is important related to ineffective breast cancer in developing countries.

Most women come to the hospital in the early stages of breast cancer and they will get surgery as the primary treatment [18]. However, the majority of Indonesian patients approximately 60%-70% are diagnosed at advanced stages between stage III and stage IV, and 35% of them have metastases [12][21][28], most Indonesian patients receive combination treatment. According to the Indonesian Society of Surgical Oncology, the breast cancer treatment consists of two types of therapy, the first is local therapy, and the second is systemic therapy. Local therapy is used to treat tumors in site without to affect the whole of the body. Surgery and radiation are both included local therapy. While systemic therapy refers to drugs which that spread to whole of the body to eliminate or suppress growth of cancer cells. Chemotherapy, hormonal therapy and molecular targeting therapy (biologic therapy) are included systemic therapy [2]. The treatment plan is depends on tumor size, differentiation's degree and axillary metastases. Each type of therapy can be applied separately or combination.

The explanations above are interesting that can be analyzed the relationship between treatment and breast cancer patients recovery. Breast cancer patient recovery in the United States of America was increased, but how about Indonesian breast cancer patient recovery, which it can be examined medical approach and several methods of data mining. The factors that influence breast cancer patients in Indonesia have different characteristics are compared with the other countries. For example, lifestyle factors tend to be a breast cancer factor in the United States of America, while the control routinely as one of breast cancer patient factor in Indonesia.

This research uses cancer registry data, which is obtained from one of the Oncology Hospital in East Java, Indonesia. We use several methods in data mining to improve the raw data, prepare clean data as training data for learning. The first step, we use preprocessing begin from data cleaning, data integration and data normalization. The second step, we use feature selection algorithm to rank attributes, which the comparison of three methods are Information Gain, Fisher's Discriminant Ratio and Chi-square [27]. The third

step, we use Hierarchical K-means Clustering to get the ideal number of attributes that can be removed with clustering analysis is error ratio and variance.Hierarchical K-means Clustering is optimization of K-means,it is able to optimize the initial centroid of K-means in several times [16]. The fourth step, we use Naive Bayesand Decision Tree as classification algorithms with Leave-one-out (LOO) as validation sampling.Furthermore, this paper will give the exploration of fact and knowledge from breast cancer data distribution.

2. RELATED WORKS

JareeThongkam,*et al* [14]from VictoriaUniversity, Australia. Their research proposed about breast cancer survivability via AdaBoost Algorithm. Their research used data mining approach to obtain the information on medical issues, improve medical checkup results, reduce treatment costs, and prioritize clinical studies of patient health. In the preprocessing was used RELIEF algorithm to select the important attributes, while to extract knowledge from database of breast cancer patient survivability using AdaBoost Algorithm. The number of patient samples was obtained from Srinagarind Hospital, Thailand, which consists of 394 patients died and 342 patients survived. There are 11 attributes or 11 categorical attributes and 2 classes in their research consist of age, marital status, occupation, basis of diagnosis, topography, morphology, extent, stage, received surgery, received radiation, received chemo, and survivability (classes). RELIEF algorithm ranked the attributes based on 7 highest scores in building the breast cancer model, including extent, stage, basis age, morphology, and occupation. The prediction accuracy results by 10-fold cross validation using Modest AdaBoost algorithm, after the feature selection is 68.63% (accuracy), 79.95% (sensitivity), and 55.70% (specificity).

Cheng-Tao Yu, *et al* [7] from National YunlinUniversity of Science and Technology, Taiwan. Their research proposed about prediction of survival in patients with breast cancer using three artificial intelligence techniques. They argue that advancement of medical technology impact the large amounts of data related with health increasingly. The prediction using data mining became an important instrument for the management of hospitals and medical research.Breast cancer dataset in their research was obtained from a regional teaching hospital in central Taiwan between 2002 and 2009. Prognostic factors of breast cancer dataset consist of 8 attributes, while the number of patient samples is 967 patients (861 samples of the patients who survived after treatment, 106 samples who died). There are two data types which were used in their research, these are categorical variable (chemotherapy, radiotherapy, and pathological, staging), and continuous variable (age, tumor size, number of lymph node examined, and number of lymph nodes attacked). The important attributes selected based on TNM (Tumor-Nodes-Metastasis) and NPI (Nottingham Prognostic Index) indicators for the prediction of survival in patients with breast cancer. The

prediction accuracy results by 10-fold cross validation using three artificial intelligence techniques are 90.31% for Artificial Neural Networks (ANNs), 89.79% for Support Vector Machine (SVM), and 88.64% for Bayesian Classifier.

Abdelghani Bellaachie, *et al* [1] from The George Washington University, Washington DC. Their research proposed about predicting breast cancer survivability using data mining techniques. They presented data mining techniques to predict the survival rates of breast cancer patients by using SEER (Surveillance Epidemiology and End Results) public data. Furthermore, their research introduced a pre-classification approach that consider in 3 variables, called Survival Time Record (STR), Vital Status Record (VSR), and Cause of Death (COD), pre-classification given result the number of patients who 76.8% survived (116.738 samples) and 23.2% not survived (35.148 samples). The number of attributes in SEER public data is 16 attributes, while the number of samples is 151.886 instances.The data type in SEER dataset consists of nominal and numerical, the nominal attributes include race, marital status, primary site node, histologic type, behavior code, grade, extension of tumor, lymph node involvement, site specific surgery code, radiation, and stage of cancer, while the numerical attributes include age, tumor size, number of positive nodes, number of nodes, and number of primaries. Their research used Information Gain (IG) to determine the contribution of each attribute, extension of tumor have the highest contribution in data. The prediction accuracy results by 10-fold cross validation are 84.5% for Naïve Bayes, 86.5% for Artificial Network, and 86.7% for C4.5.

R. K. Kavitha, *et al* [23] from Vinayaka Missions University, Tamil Nadu. Their research proposed about predicting breast cancer survivability using Naïve Bayesian classifier and C4.5 algorithm. They analyzed SEER public data which it is pre-classified to make decision about prognosis of breast cancer. Their research used a preprocessing SEER data to select parameters which are not related with breast cancer such as race, ethnic, and all related social demographics. SEER data has 124 attributes which they were reduced to be 5 attributes only, begin from removed the attributes that contains social demographics, missing values, duplicate, same values, and the final process was obtained 1.153 selected samples of 1.403 samples without missing values. The selected attributes after preprocessing are age, clump thickness, menopause, tumour size, and CS extension. The prediction accuracy results by 10-fold cross validation are 95.79% for Naïve Bayes and 97.7% for C4.5.

HadiLotfnezhadAfshar, *et al* [10] from University of Medical Sciences, Tehran, Iran. Their research proposed about prediction of breast cancer through knowledge discovery in databases. They argue that current medical data collection is very large, it gives an opportunity for researcher in the world to develop a predictive model of patient survivability through the medical research community. Their research developed a prediction model and discovered the relationship between predictor variable and survival of

breast cancer patient variable. The data was used in this research from SEER public data that has 72 attributes and 657.712 samples, but preprocessing attributes or variables that selected as many as 18 important attributes which it removed the attributes that not related with breast cancer factors. The data type was used in their research dataset consists of categorical and continuous, the categorical attributes include race, marital status, primary site node, histology, behavior, grade, extension of tumor, lymph node involvement, radiation, stage, site specific surgery code, ER status, and PR status, while the continuous attributes include age, tumor size, number of positive nodes, number of nodes, and number of primaries.The relative importance of predictor variables are identified by SVM, they are behavior, lymph node involvement, extension of tumor, grade, number of positive nodes, age, site specific surgery code, PR status, radiation, stage, and other. The prediction accuracy results are 86.7% for SVM, 83.9% for Bayes Net, and 82.4% for CHAID (Chi-squared Automatic Interaction Detection).

3. ORIGINALITY

Breast cancer is the second highest cancer type which attacked Indonesian women [25]. The adherence of Indonesian society within breast cancer control to the hospital becomes a very important consideration related with the patient's recovery. Therefore, we hope the data mining techniques can produce the better data model and useful information.This paper proposes the different research flow stepwith previous related works on the breast cancer dataset problems, by combining the technique: (1) Preprocessing usingdata cleaning, data integration, and data normalization (2) Feature selection using Information Gain, Fisher's Discriminant Ratio, and Chi-square (3) Remove the lower contribution attributes using Hierarchical K-means clustering (4) Build the classification model using Naive Bayes, and Decision Tree (5) Validation sampling using leave-one-out which it will predict the breast cancer class to per sampledefined as per patient. As for the result, we will get the most important attributes from the existing attributes that involved in data by the comparison of three feature selection algorithms. The performance evaluation of classification model is focus on precision which it evaluates the patient who true positive is predictedthathas breast cancer, the patient who false positive is predictedthathas breast cancer. Furthermore, breast cancer dataset will be analyzed to find the interesting facts from data distribution about the relationship among two or more variables such asthe relationship of treatment between patient who evidence of cancer (not recover yet), or no evidence of cancer (recover).

4. SYSTEM DESIGN

The proposed system consists of 5 phases: (1) Data collection, (2) Data preprocessing, (3) Attribute ranking and determination of removal, (4) Classification model and evaluation, (5) Output prediction. The whole system

design is shown in Figure 2. Each phase on system design will be explained in part 4.1-4.5.

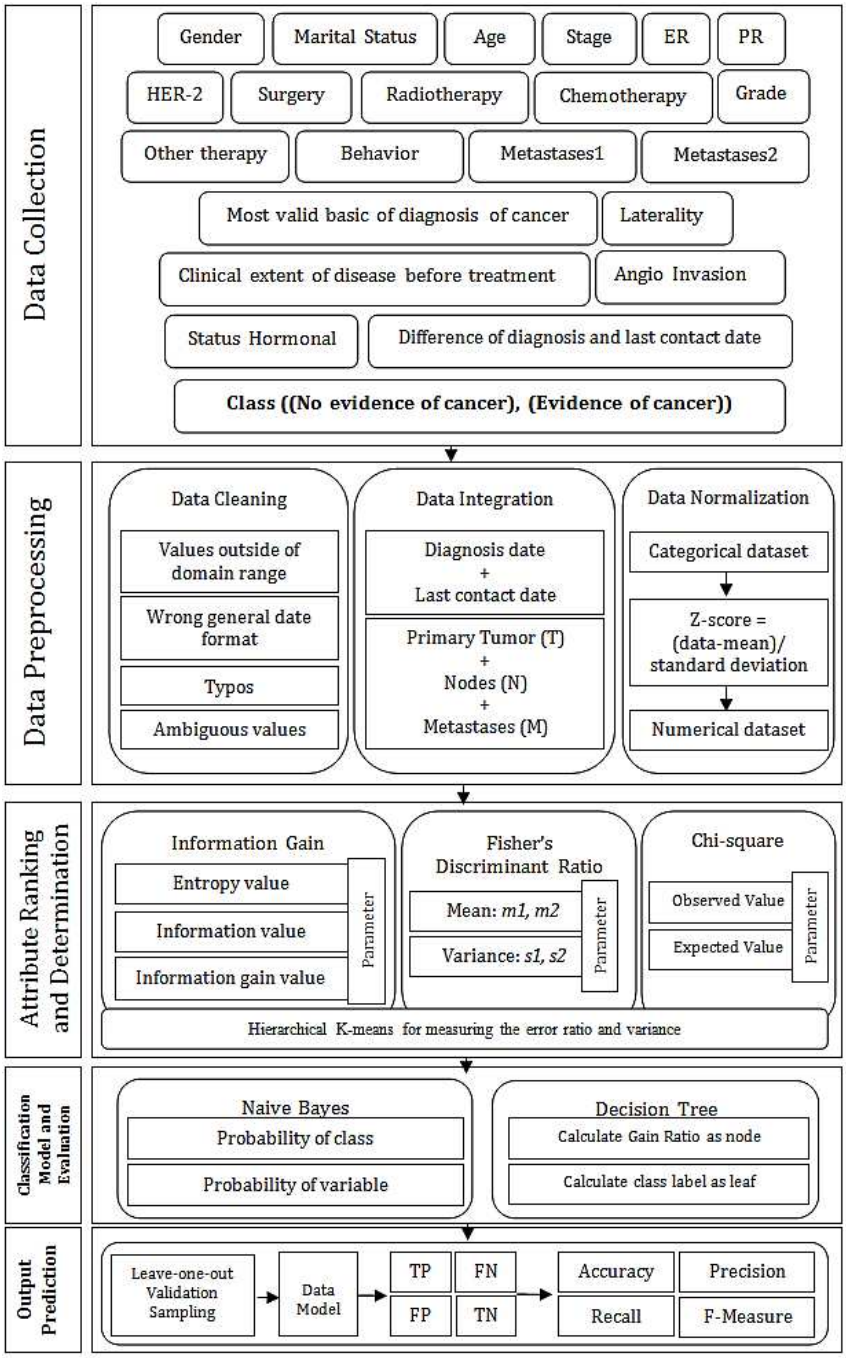

Figure 2. System design of our proposed approach for breast cancer research

4.1. Data Collection

The proposed system uses original sample of oncology hospital patients. Breast cancer dataset consistsof 21 attributes, 1097 samples, two classes, and data was taken from the last 3 years (2009-2011) in this research.The other

breast cancer dataset generally has two classes of patient, which consist of normal and abnormal, or survive and failure. But, our dataset is different with previous research, because our dataset consists of all patients who have breast cancer, and then provides the patient information is classified as recover or not yet after the patient was given treatment. Furthermore, our dataset represents the sample distribution of breast cancer patient especially in East Java, Indonesia which has different with other dataset and analysis result.

4.2. Data Preprocessing

The data preprocessing is preparing data to be a fixeddata, before the data will be a training data.This task depends on the data mining expert for improving the data quality, increasing accuracy and effectiveness of data mining process. The preprocessing task will take 60% effort of data mining process [29]. We are following the major tasks in data preprocessing that consist of data cleaning, data integration, data reduction, and data transformation and discretization.

4.3. Attribute Ranking and Determination of Removal

Feature selection is used to search the features or attributes that have contributed or more weights in dataset, it is the process of selecting a subset of relevant features for building powerful model. The poor classification results are usually affected by features that have small contribution. Feature selection is used for (1) simplifying the models in order to make easier to interpret by researcher or users, (2) reducing the time-consuming of training data, (3) increasing generalization with reducing overfitting (variance reduction).Feature selection also useful toremove the irrelevant and redundant features, reduce the computation cost, and provide the relevant data selection [13][26].

The space of characteristic of feature selection algorithm consists of 3 categories: search organization, generation of successor states, and evaluation measures [17]. We proposed to use the evaluation measure on our breast cancer dataset and to get the ranking of each feature which it is helpful the doctor or medical experts to know the most important features of breast cancer dataset.We are using evaluation measures that consist of information, dependence, and divergence criteria in our research which the feature selection algorithm is represent those criteria. Evaluation measure is function to evaluate successor candidate which comparing different hypothesis to advise the search process. Our research is using the comparison of three feature selection algorithms such as (1) Information Gain represents potential information measurement criteria, (2) Fisher's Discriminant Ratio represents divergence measurement criteria, and (3) Chi-square represents dependence measurement criteria. The result of three feature selection algorithms using those criteria will be validated by medical experts.

4.3.1 Information Gain Concept

Information theory, Entropy, and Information Gain are several basic concepts invented by Claude Shannon in 1948 [24]. The data measurement required impurity levels in the sample group. An impurity examination needs to determine the data quality. The impurity levels in data are illustrated in Figure 3 below.

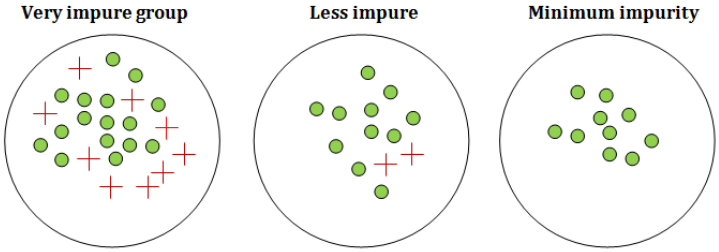

Figure 3. The impurity levels in data group

The impurity measurement can use entropy formula which the uncertainty size is associated with the random variable. The calculation of discrete random *Y* variable take *m* different value $\{y_1, y_2, ..., y_m\}$, the entropy formula as follows below:

$$H(Y) = -\sum_{i=1}^{m} p_i \log(p_i), \quad where \quad p_i = P(Y = y_i) \tag{1}$$

$p_i$is the probability of class *i* which calculates the each class *i* proportion in the set. The entropy of two possibilities cases with probabilities *p* and *q=1-p.*

$$H = -(p \log p + q \log q) \tag{2}$$

Our research is using Information Gain to get the ranking of features from top-bottom ranking with calculating the entropy, information, and information gain value.

4.3.2 Fisher's Discriminant Ratio Concept

Discriminant function analysis or the original dichotomous discriminant analysis was developed by Sir Ronald Fisher in 1936.This method is a statistical analysis to predict a categorical dependent variable, and more continuous or binary independent variables.Fisher's Discriminant Ratio (FDR) is commonly used to measure the strength of discrimination the individual features in separate two classes based on its value, the process of splitting class is illustrated in Figure 4. The expression of m_1 and $m_2$are the average value of two classesrespectively, while s^2_1 and $s^2_2$are variance of the two classes in the feature to be measured respectively. FDR formula is defined as following equation:

$$FDR = \frac{(m_1 - m_2)^2}{(s^2_1 + s^2_2)} \tag{3}$$

Where :
FDR = Fisher's Discriminant Ratio
m_1 = *mean of class 1*
m_2 = *mean of class 2*
s_1 = *variance of class 1*
s_2 = *variance of class 2*

The result is given by FDR are the features that have a maximum difference in the average of the class and minimum variance of each class, therefore the high FDR value will be obtained.If two features have the average absolute difference that is equal but the number of variance is different, therefore the feature with the minimum number of variance will get the high FDR value. In other hand, if two features have the number of variance that is equal, but the average absolute difference is different, therefore the feature with the maximum average absolute difference that will get the high FDR value [9].

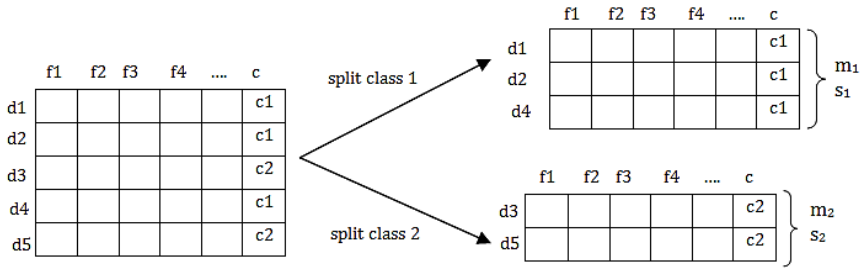

Figure 4. The process of splitting class on FDR

Our research is using FDR to get the ranking of features from top-bottom ranking with splitting two classes, after that calculating the mean and variance of each class.

4.3.3 Chi-square Concept

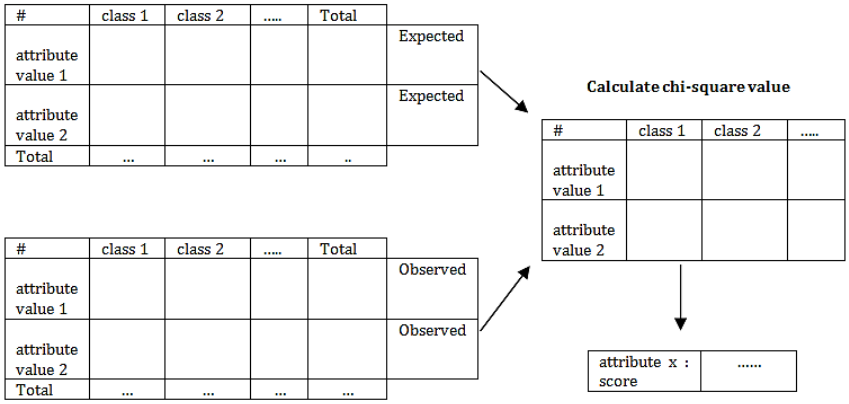

Figure 5. The process of calculating expected and observed value on Chi-square

Chi-square (X^2) is a statistical method that is applied to test independent of two events. Chi-square feature selection is used to evaluate attribute value with calculate chi-square value related to class [27]. Chi-square calculate the sampling distribution of each feature is defined as expected value, while calculate the total data of each class, total data of each class attribute, and total data of all classes is defined as observed value. The calculation of expected and observed value is used to get chi-square value for each class attribute.

The total of chi-square value will be accumulated to final score of each attribute, the process of calculating expected and observed value is illustrated in Figure 5. Chi-square formula is defined as following equation:

$$X^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \quad (4)$$

Where :

X^2 = *Chi-square*
$\sum$ = *the sum of total chi-square*
O = *Observed score*
E = *Expected score*

Our research is using Chi-square to get the ranking of features from top-bottom ranking with calculating the sampling data distribution, expected values, and observed values.

4.3.4 Attribute Determination of Removal using Hierarchical K-means

Our research is using clustering to obtain error ratio and variance from grouping data naturally (without label) and validate the clustering results with the original label that is already available class from supervised data, the error ratio and variance is generated later used to consider removing the ideal number of attributes from ranking of feature selection results. We are using clustering for removing the lower contribution attributes, because the measurement of how good data can be proved using clustering before we implement classification algorithms (to build the classification model). Clustering measures the data which has some similarities characteristic will gather in the same cluster, and data which has different characteristics will gather in the different cluster. It means clustering can separate the patient that belongs to cancerclass label (evidence of cancer) and recover class label (no evidence of cancer) based on parameter measurement of error ratio and variance.

This research used Hierarchical K-means Clustering [16], which this method is an optimization of the K-means before. This method is able to handle the K-means clustering problems that often reach local optima. Hierarchical K-means can improve the better cluster results, because it is able to optimize the initial centroid of K-means several times. This algorithm transforms all the centroids of clustering with combine Hierarchical Clustering to determine the initial centroids for K-means. This algorithm is

better used for clustering cases that complex with large dataset and high number of dimensions. Hierarchical K-means offers the advantage in the speed side by K-means algorithm and the precision side by Hierarchical Algorithm.

The measurement of clustering analysis uses the error ratio and variance. Error ratio is used to determine the number of data misclassified and the total number of data.

$$Error\ ratio = \frac{misclassified}{totaldata} \quad x100\% \tag{5}$$

Variance is used to determine Vw and Vb, which Vw is variance within clusters and Vb is variance between clusters. Ideal cluster has internal homogeneity expressed by minimum variance within cluster (Vw) and external homogeneity expressed by maximum variance between clusters (Vb).

$$V = \frac{V_w}{V_b} \quad x100\% \tag{6}$$

Our research applied this algorithm to remove the lower contribution attributes, one by one feature is removed from data based on the ranking features of three feature selection algorithm and then we measures using Hierarchical K-means to get the ideal error ratio and variance.

4.4. Classification Model and Evaluation

Classification is performed directly from the relationship of training data to testing data. Classification is close with concerned prediction which this method builds a model called predictive modeling that can perform mapping of each set of variables to the class target, thereafter use the model to provide a target value on the new set that obtained. Classification algorithm typically consists of two phases:

- Training phase: a model is constructed from the training data.
- Testing phase: the model is used to assign a label to an unlabeled testing data.

The classification flow of building model is illustrated in Figure 6 below:

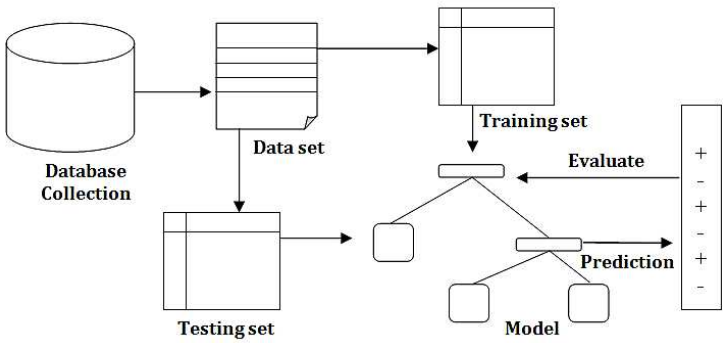

Figure 6. Classification of building model flow

The classification algorithm based on training method is divided into two types: eager learner and lazy learner. In our research is using Naïve Bayes and Decision Tree which they are eager learner. The eager learner is designed for learning the training data to map each input vector to class label, at the final training process the model can already be mapped correctly all training data to class label.The advantage of eager learner method is running prediction process quickly, but must be paid to the long training process. To support the measurement of our breast cancer dataset which has categorical data, then we applied Naïve Bayes and Decision Tree classifiers to build the classification model. We considered that Naïve Bayes classifier is appropriate to solve the categorical data model using probabilistic method, while Decision Tree classifier is appropriate to solve the categorical data model using decision making cases like visualization of tree which it is represented by reasoning procedures.

4.4.1 Naive Bayes

Probabilistic method is the most fundamental of all data classification methods [6]. Probabilistic classification method uses statistical conclusion to find the best class for a given example. The popular classification method of probabilistic is Naive Bayes classifiers which is asimple probabilistic classifiers family based on applying Bayes theorem with strong independence assumption between the features. Thomas Bayes (1702-1761) is who proposed the Bayes theorem. Naive Bayes classifiers also represents a supervised learning method as well as a statistical method for classification, assume a probabilistic model that underlie and enable to capture uncertainty about the model in a principled way to determine the outcome probability. It can solve the problem of diagnostic and predictive.

The Naïve Bayes theorem explanation, note that the classification process requires a number of clues to determine what classes are suitable for the sample analyzed. Therefore, the Bayes theorem as follows [5]:

$$P(C \mid F_1 \ldots F_n) = \frac{P(C)P(F_1 \ldots\ldots F_n \mid C)}{P(F_1 \ldots F_n)} \tag{7}$$

Where the C variable represents the class, while the $F_1 \ldots\ldots F_n$variable represents the characteristic of the directions that needed to the classification.The formula above (7) it can be concluded that the naïve independence assumption makes conditional probability to be simple, therefore the calculation becomes possible to do. The next steps, the $P(C|F_1 \ldots F_n)$ formula can be simplified to:

$$\begin{aligned} P(C \mid F_1 \ldots F_n) &= P(C)P(F_1 \mid C)P(F_2 \mid C)P(F_3 \mid C)\ldots \\ &= P(C)\prod_{i=1}^{n} = P(F_i \mid C) \end{aligned} \tag{8}$$

Our research is using Naïve Bayes classifiers in categorical models which the calculation involves probability of each variable and class.

4.4.2 Decision Tree

Decision Tree is trees that used as reasoning procedures to get the answer from the problem which is entered. The flexibility of decision tree makes this method attractive, especially because it gives the advantage of advice visualization (as decision tree) to make prediction procedure can be observed [9]. Decision tree is most used to solve decision-making cases such as medical field (e.g. diagnosis of patient disease), computer science (e.g. data structure), psychology (e.g. decision-making theory), and etc.

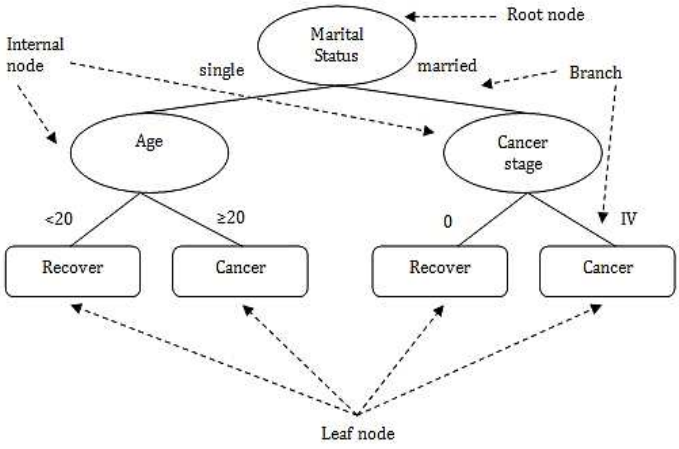

Figure 7. The example of decision tree visualization

The characteristic of decision tree is illustrated in Figure 7, which there are elements as follows:

- **Root node,** has not input branch but has more output branch.
- **Internal node,** each node that not a leaf which has one input branch and two or more output branch. This node express of testing is based on feature value.
- **Branch,**each branch express of testing result value in node which is not a leaf.
- **Leaf node,**node has one input branch exactly and has not output branch. This node express of the class label.

The long or short rule that is generated depends on type of decision tree algorithm which is used. There are two popular type of decision tree which is often used by researcher, such as ID3 and C4.5. ID3 is using Entropy, Information Entropy, and Information Gain. C4.5 or Classification version 4.5 is the development of ID3 algorithm which C4.5 algorithm has a same basic principle of ID3 algorithm. The main difference between C4.5 with previous version is:

- C4.5 can handle the continuous and discrete attributes.
- C4.5 can handle the training data which has missing values.
- C4.5 has pruning process of decision tree
- The selection of attributes using Gain Ratio.

C4.5 is the successor of ID3 which uses Gain Ratio to improve Information Gain formula:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \quad (9)$$

where:

S	*=sample space (data) that is used for training*
A	*=attributes*
Gain(S,A)	*= Information Gain to A attribute*
SplitInfo (S,A)	*= Split Information to A attribute*

Attribute that has highest Gain Ratio value selected as test attribute to the node. With Gain using Information Gain, this approach apply normalization of Information Gain using Split Information. SplitInfo express the entropy or potential information by the formula:

$$SpitInfo(S, A) = -\sum_{i=1}^{k} \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (10)$$

where:

S	*= sample space (data) that is used for training*
A	*=attributes*
S_i	*= the number of sample to attribute-i*

Our research is using C4.5 of decision tree algorithm which this algorithm can generate the short decision tree model, easy to understand that model, and give the better prediction result.

4.5. Output Prediction

The validation sampling is a process to divide between training and testing data before the classification algorithm build the model. The current research is often using several validation models such as Holdout, Random subsampling, K-fold cross validation, Leave-one-out (LOO) cross validation, and Bootstrap. Our research is using Leave-one-out cross validation which *K* is chosen as the total number of examples (LOO is the generated case of K-fold cross validation).

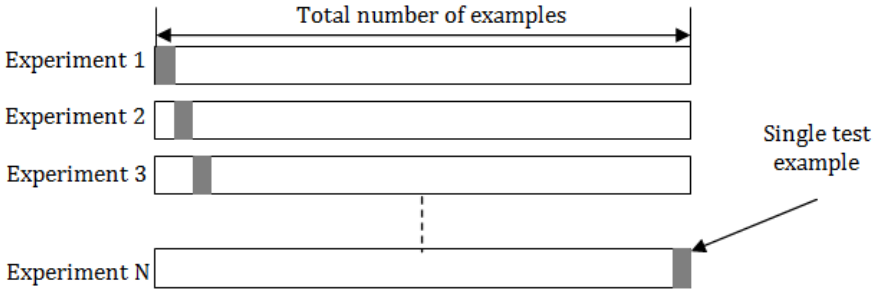

Figure 8. Leave-one-out cross validation flow

Our research chooses LOO cross validation, because we uses medical dataset that contain each sample as individual patient data, and it needs to validate one by one as testing data that will be predicted. Figure 8 is illustrated Leave-one-out cross validation flow.

$$E = \frac{1}{N}\sum_{i=1}^{N} E_i \quad (11)$$

where:
E = *Experiment*
E_i =*the number of experiment-i*
N =*the number of example*

Classification performance evaluation could be calculated using measurement such as accuracy, precision, recall, and F-measure which is described in Table 1. Our research focus on precision that is used to evaluate patients who are predicted as evidence of cancer (not recover yet) and who are false predicted as evidence of cancer.

Table 1. Classificationperformance evaluation

Measurement	**Formula**
Accuracy	TP+TN/ TP+TN+FP+FN
Precision	TP / TP+FP
Recall	TP/TP+FN
F-Measure	2*Precision*Recall / Precision + Recall

5. EXPERIMENT AND ANALYSIS

This chapter explainshow measure and analyze the breast cancer patient data using data mining approach, which it aims to determine the effectiveness of the recovery process of breast cancer patient by the specific experiment: (1) Breast cancer dataset, (2) Preparing dataset, (3) The comparison of three feature selection algorithms, (4) The comparison of Naive Bayes and Decision Tree algorithms, and (5) The interesting facts and analysis of breast cancer dataset.

5.1. Breast Cancer Dataset

Breast cancer dataset in this research was obtained from cancer registry of Oncology Hospital from 2009-2011. This data source is the form of raw data that have the patient's identity, such as registration number, name, religion, ethnic and city. The attributes were included in this research related with factors of breast cancer patient diagnosis results. The attributes are categorical type, which the variation of the value of each attributeismost representative and informative.

5.2. Preparing Dataset

Data preprocessing is an important stage in data mining, because it can be handle various types of dirty data on large datasets. The selection of

appropriate data preprocessing methods with certain issues is very important. This role depends on expert data mining to improve data quality and increase the accuracy. The form of preprocessing in our research is data cleaning, data integration, data reduction and data transformation.

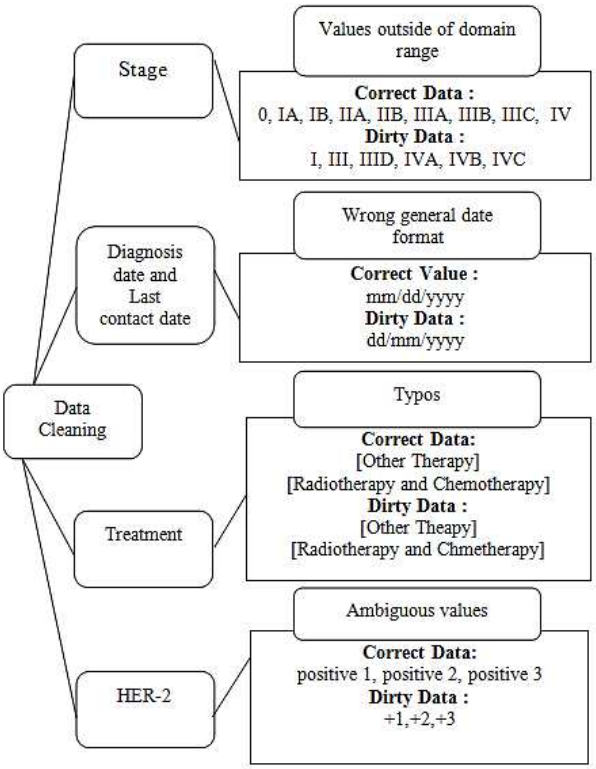

Figure 9. Data cleaning process of breast cancer dataset

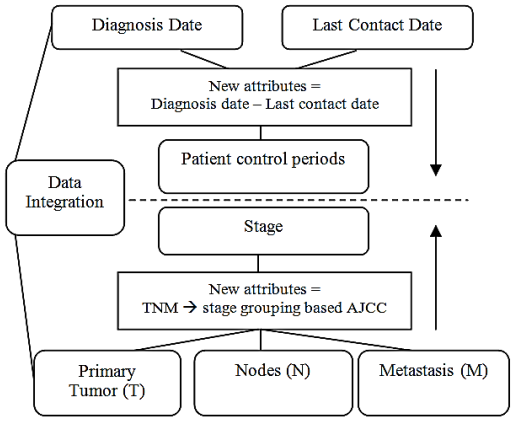

Figure 10. Data integration process of breast cancer dataset

5.2.1 Data Cleaning

Data cleaning is a technique to find and repair the dirty data, corrupt, and inaccurate. Breast cancer dataset in this research is the raw data that obtained every year, should be preprocessed and improved to become a dataset that can be used to a learning model. Data cleaning process is following the steps below:

- Remove features that are related to the patient's identity, such as registration number, name, religion, ethnic and city.
- Improve data that qualify as dirty data, Figure 9isshown the data cleaning process begin from detect breast cancer dataset which are value outside of domain range, general wrong date format, typos, and ambiguous values (it is numeric or string).

5.2.2 Data Integration

Data integration is a technique for combining two or more attributes from various sources and provides standardization on the value of each attribute. Data integration in our research was applied on diagnosis date and

last contact date attributes which it was merged to one attribute as patient control period attribute, while stage attribute was obtained from T-N-M (Primary Tumors-Nodes-Metastasis) that has been grouped by AJCC (American Joint Committee on Cancer) 2010. Data integration on breast cancer dataset is shown in Figure 10.

5.2.3 Data Transformation

Data transformation is a technique to change the data in a certain range, one of data transformation technique is normalization which is the process of scaling data to provide a range of values that is balanced in each dimension data. Normalization that was used in this research is Z-score, which it calculates based on mean and standard deviation of data. Z-score formula as follows:

$$new\_data = \frac{data - mean}{std} \quad (12)$$

Where :

std = *standard deviation*

This research used categorical data type, which data has been processed in data cleaning and data integration with 1097 samples. In order to categorical data can be processed into normalization, it must be converted to numerical. The purpose of normalization in this breast cancer dataset would be processed inHierarchical K-means clustering which it required numerical data.

Table 2. Convert categorical data to Z-score

Gender	***features***		***Gender***	***Features***		***Gender***	***features***
male	[category1]		1	[numeric1]		0.042737	[z-score1]
female	[category2]	*convert to*	2	[numeric2]	*convert to*	2.611413	[z-score2]

The process of converting a value of categoricalattribute to numerical and then convert to Z-score is shown on Table 2. Z-score data that will be a training data which is prepared for clustering, while categorical data is prepared for feature selection and classification.

5.3. The Comparison of Three Feature Selection Algorithms

Our research applied three feature selection methods that consist of Information Gain, Fisher's Discriminant Ratio, and Chi-square. Those output of three algorithms given ranking of each attribute which it could be visualized from the highest and lowest contribution attributes.

5.3.1 Information Gain

Information Gain is better applied to measure the attributes or features on supervised data independently, therefore that each attribute will be

measured by Information Gain. In chapter 4.3.1 was explained the conceptual of Information Gain.The Information Gain formula as follows:

- Calculate the entropy values
  $Info\ (D) = -\sum^{m}_{i=1} pi\ log_2\ (pi)$ (13)
- Calculate information values
  $Info_A(D) = \sum^{y}_{j=1} \frac{|\ Dj\ |}{|\ D\ |} x\ Info\ (D_j)$ (14)
- Calculate information gain values
  $Gain\ (A) = Info\ (D)\text{-}Info_A(D)$ (15)

Table 3. Information Gain result of breast cancer dataset

Rank.	Information Gain	
1	Surgery	0.5234999
2	Angio Invasion	0.3737964
3	Grade	0.3221062
4	Difference of diagnosis and last contact date	0.2954116
5	Most valid basic of diagnosis of cancer	0.268414
6	ER	0.2315164
7	PR	0.2274813
8	HER-2	0.1944091
9	Clinical of extent diseases before treatment	0.1614221
10	Chemotherapy	0.1454332
11	Site of distant Metastases 1	0.1312649
12	Stage	0.0745327
13	Site of distant Metastases 2	0.0347847
14	Radiotherapy	0.0283093
15	Behavior	0.0277904
16	Age	0.0111039
17	Laterality	0.0055416
18	Other therapy	0.0042015
19	Hormonal status	0.001335
20	Marital status	0.00061
21	Gender	0.0000161

The calculation result of Information Gain on breast cancer dataset is shown in Table 3.The calculation of Information Gain on each attributeis applied recursively. The Table 3above described that the surgery attribute has highest Information Gain score (0.5234999), while chemotherapy and radiotherapy attributes have Information Gain scores(0.1454332 and 0.0283093 respectively), both are in the middle ranking. The other therapy attribute is bottom ranking (0.0042015), while gender attribute is lowest

Information Gain score (0.0000161). The following Figure 11 below is shown a chart of attribute ranking using Information Gain.

5.3.2 Fisher's Discriminant Ratio

Table 4. Fisher's Discriminant Ratio result of breast cancer dataset

Rank.	FDR	
1	Surgery	3.532512026
2	Angio Invasion	1.457139212
3	Grade	1.316148239
4	Difference of diagnosis and last contact date	0.870841069
5	ER	0.742281934
6	PR	0.688065472
7	Most valid basic of diagnosis of cancer	0.617575637
8	Chemotherapy	0.489646753
9	HER-2	0.28837049
10	Clinical of extent diseases before treatment	0.244711231
11	Site of distant Metastases 1	0.209935732
12	Stage	0.093935953
13	Radiotherapy	0.07984221
14	Site of distant Metastases 2	0.043038236
15	Behavior	0.042362973
16	Other therapy	0.011317328
17	Laterality	0.007013754
18	Hormonal status	0.002591769
19	Age	0.002300267
20	Marital status	7.98E-05
21	Gender	4.49E-05

Fisher's Discriminant Ratio (FDR) is commonly used to measure the strength of discrimination the individual features in separate two classes based on its value. In chapter 4.3.2 was explained the conceptual of FDR.The Fisher's Discriminant Ratio in our breast cancer research followed the algorithm below:

- Split data, which data is separated to two classes between 'No evidence of cancer' as class 1 and 'Evidence of cancer' as class 2.
- Calculate the total of each individual data and mean of data for each class 1 and class 2, which the number of class 1 is 488 samples and class 2 is 609 samples, therefore it is obtained m_1 and $m_2$values for each attribute.
- Calculate the variance of each class using m_1 and m_2 values that has been obtained, which the variance is $(data\text{-}mean)^2$divided by number of populationdata for each class, therefore it is obtained s^2_1 and s^2_2 valuesfor each attribute.

- Calculate the FDR value for each attribute.

The calculation result of Fisher's Discriminant Ratio on breast cancer dataset is shown in Table 4.The Table 4 above described that the surgery attribute has highest Fisher's Discriminant Ratio score(3.532512026), while chemotherapy and radiotherapy attributes have Fisher's Discriminant Ratio scores(0.489646753 and 0.07984221 respectively), both are in the middle ranking. The other therapy attribute is bottom ranking (0.011317328), while gender attribute is lowest Fisher's Discriminant Ratio score (0.0000161). The following Figure 12 below is shown a chart of attribute ranking using Fisher's Discriminant Ratio.

5.3.3 Chi-square

Table 5. Chi-square result of breast cancer dataset

Rank.	**Chi-square**	
1	Surgery	694.2289
2	Angio Invasion	521.0985
3	Grade	427.893
4	Difference of diagnosis and last contact date	405.8851
5	Most valid basic of diagnosis of cancer	343.0246
6	ER	313.3043
7	PR	308.852
8	HER-2	276.2856
9	Chemotherapy	213.9797
10	Clinical of extent diseases before treatment	197.1965
11	Site of distant Metastases 1	147.5819
12	Stage	104.1069
13	Radiotherapy	43.0666
14	Site of distant Metastases 2	39.8705
15	Behavior	33.6873
16	Age	16.0202
17	Laterality	6.7775
18	Other therapy	5.7345
19	Hormonal status	2.0302
20	Marital status	0.9258
21	Gender	0.0247

Chi-square (X2) is a statistical method that was applied to test independent of two events. Chi-square feature selection is used to evaluate attribute value with calculate Chi-square value related to class. In chapter 4.3.3 was explained the conceptual of Chi-square. The Chi-square in our breast cancer research followed the algorithm below:

- Calculate the sampling distribution of each feature is defined as expected value.
- Calculate the total data of each class, total data of each class attribute, and total data of all classes is defined as observed value.
- Calculate chi-square value for each class attribute from expected and observed value.
- The total of chi-square value will be accumulated to final score of each attribute.

The following Table 5 is shown the calculation results of Chi-square on breast cancer dataset in this research.The Table 5above described that the surgery attribute has highest Chi-square score (694.2289), while chemotherapy and radiotherapy attributes have Chi-square scores (213.9797 and 43.0666 respectively), both are in the middle ranking. The other therapy attribute is bottom ranking (5.7345), while gender attribute is lowest Chi-square score(0.0247). The following Figure 13 below is shown a chart of attribute ranking using Chi-square.

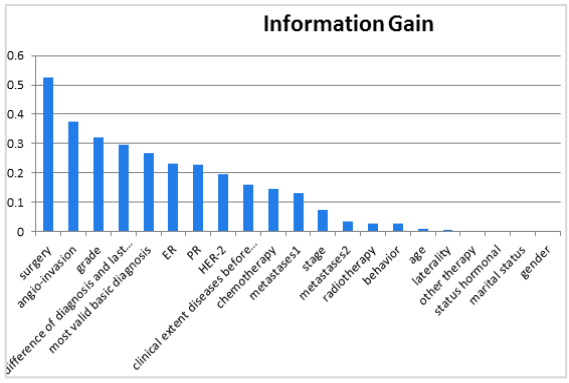

Figure 11. Attribute ranking using Information Gain of breast cancer dataset

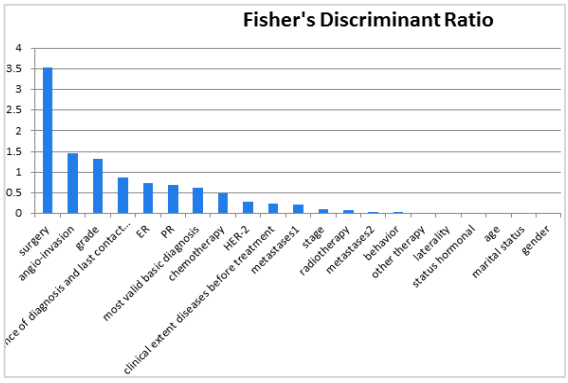

Figure 12. Attribute ranking using FDR of breast cancer dataset

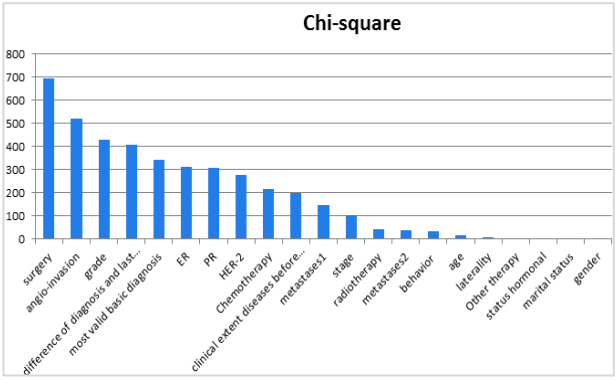

Figure 13. Attribute ranking using Chi-square of breast cancer dataset

5.3.4 The Removal of Low Contribution Attributes

The feature selection result from three different methods was obtained. Furthermore, how to remove the ideal number of attributes that have low contribution attributes. We made a rule to remove the attribute that has low contribution in data. We called as *fs (feature selection)-$_n$* ,where *n* represented the number of attributes that are removed. For example, attribute ranking based on feature selection results by Information Gain, fs_1 is represented to remove gender attribute, $fs_2$is represented to remove gender and marital status, $fs_3$is represented to remove gender, marital status and hormonal status, and until fs_n which there are only two attributes that are left out to be processed in Hierarchical K-means, therefore is obtained error ratio and variance.

How to choose the ideal number of attributes that can be removed from data, Table 6 below can be a solution to get the difference of each shifting *fs* from error ratio and variance with comparison *fs* on feature selection results.The following calculation formula below:

$$fs_{(n),(n+1)} = fs_{(n)} - fs_{(n+1)} \quad (16)$$

Where :

$fs_{(n)}$ = *The error ratio and variance values on current feature removal*

$fs_{(n+1)}$ = *The error ratio and variance values on next feature removal*

Table 6. The calculation of shifting*fs* in Information Gain based on Hierarchical K-means Clustering

	fs	**Error Ratio**	**Variance**	**Difference of each error ratio**	**Difference of each variance**
	fs_1	43.75569736	0.00348199		
$fs_{1,2}$	fs_2	43.75569736	0.003306977	0	0.000175013
	fs_2	43.75569736	0.003306977		
$fs_{2,3}$	fs_3	43.66453965	0.003278214	0.09115771	2.8763E-05
	….	…….	…….		
…..	….	…….	…….	…….	…….
	fs_6	**43.75569736**	**0.002873015**		
$fs_{6,7}$	**fs_7**	**18.32269827**	**0.001965177**	**25.43299909**	**0.000907838**
	….	…….	…….		
…..	….	…….	…….	…….	…….
	fs_{18}	11.30355515	0.00045146		
$fs_{18,19}$	fs_{19}	14.85870556	3.63E-04	-3.55515041	8.85E-05

The following in Table 6, Hierarchical K-means result selected $fs_{6,7}$ which difference of error ratio fs_6 (43.75569736) and fs_7(18.32269827) is 25.43299909, while difference of variance fs_6 (0.002873015) and

fs_7(0.001965177) is 0.000907838. The difference of each error ratio and each variance of $fs_{6,7}$ is higher than other fsin this experiment. Therefore $fs_{6,7}$ selected as the ideal number of attributes can be removed that is 7 attributes.

The following Figure 14is shown the comparison among three feature selection methods, which the three methods that have highest difference of error ratio on$fs_{6,7}$ reached 25.43299909 respectively. While Figure 15 is shown the comparison of variance among three feature selection methods, which the three methods that have highest difference of variance on $fs_{6,7}$reachedInformation Gain0.000907838, Fisher's Discriminant Ratio 0.000905212, and Chi-square 0.000907838.

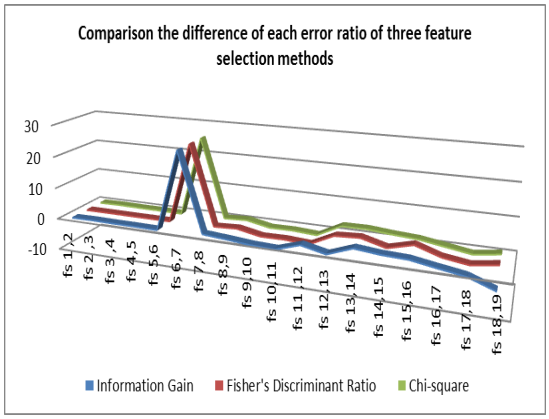

Figure 14. The comparison of difference of each error ratio of three feature selection methods

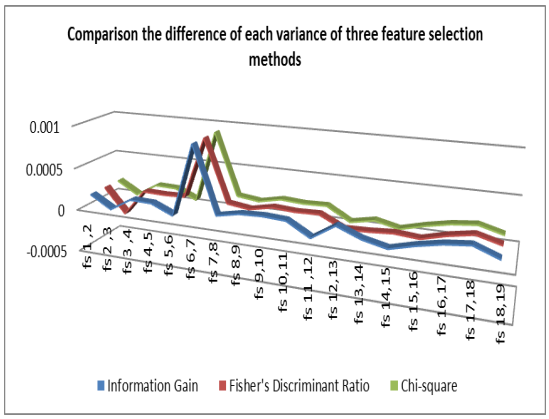

Figure 15. The comparison of difference of each variance of three feature selection methods

The selection $fs_{6,7}$ shown attribute ranking that the low contribution from feature 15 until 21 must be removed. The following Table 7 shown 7 attributes have been removed by considered the comparison of three feature selection based on error ratio and variance using Hierarchical K-means.

Table 7. The attributes were removed by three feature selection methodsbased on Hierarchical K-means Clustering

Features from low-ranking	Information Gain	Fisher's Discriminant Ratio	Chi-square	Attribute Information
feature 15	behavior	behavior	behavior	removed
feature 16	age	other therapy	age	removed
feature 17	laterality	laterality	laterality	removed
feature 18	other therapy	hormonalstatus	other therapy	removed
feature 19	hormonalstatus	age	hormonal status	removed
feature 20	marital status	marital status	marital status	removed
feature 21	gender	Gender	gender	removed

With the removed of 7 attributes which have low contribution, it decreased the error ratio of breast cancer dataset from 44.48% to 18.32%, or the accuracy increased from 55.52% to 81.68% based on clustering results.To validate of feature selection result, we also clarify to medical expert about it. The medical expert validated that 7 features (behavior, age, laterality, other therapy, hormonalstatus, marital status, and gender) can be removed from breast cancer dataset, because it is not too high influential to be involved in dataset.

5.4. The Comparison of Naive Bayes and Decision Tree Algorithms

The 14 highest contribution attributes were selected by combination between the comparison of three feature selection algorithm and the measurement error ratio and variance of Hierarchical K-means Clustering. The next step, our research appliedclassification methods that consist of Naive Bayes and Decision Tree.The final result of classification performance is accuracy, precision, recall, and F-measure of classification model.

5.4.1 Naive Bayes

Naive Bayes is supervised learning that used probabilistic model, this algorithm appropriates to solve our research about diagnostic and predictive of breast cancer. In chapter 4.4.1 was explained the conceptual of Naïve Bayes. The algorithm of Naive Bayes in our research as follows:

- Step I : Calculate the sample number of each class
  P (Cancer = No evidence of cancer) →Class 1
  P (Cancer = Evidence of cancer) → Class 2
- Step II : Calculate the sample number of each variable group on same class
  P (Gender = L | Cancer = No evidence of cancer)
  P (Gender = L | Cancer = Evidence of cancer)
  P (Gender = P | Cancer = No evidence of cancer)
  P (Gender = P | Cancer = Evidence of cancer)
  P *(…..attributes…… = ……attribute values….. | class = ….class values....)*
- Step III : Multiplication of the number of each variable on 'No evidence of cancer' and 'Evidence of cancer' class
  for example:
  (The total number of samples)=1097,
  (The number of samples | No evidence of cancer) = 488,
  (The number of samples | Evidence of cancer) = 609.

  P ((Gender = L), (... = ...), (... = ...) | Cancer = No evidence of cancer))
  = 1/488 x ... x ... x 488/1097
  P ((Gender = L), (... = ...), (... = ...) | Cancer = Evidence of cancer))
  = 1/609 x ... x ... x 609/1097
- Step IV : Compare the calculation result of class 1 and class 2
  If class 1> class 2 then class= 'No evidence of cancer'

If class 1 < class 2 then class='Evidence of cancer'

The performance metric begin from TP, TN, FP, and FN of breast cancer dataset using Leave-one-out cross validation and Naïve Bayes as follows in Table 8:

Table 8.TP, TN, FP, and FN results of breast cancer dataset using Naïve Bayes

True Positive	True Negative	False Positive	False Negative
525	447	41	84

The information of explanation in Table 8 above is:

- **True Positive**, there are 525 patients who they are predicted breast cancer correctly.
- **True Negative,** there are 447 patients who they are predicted recover correctly.
- **False Positive,** there are 41 patients who they are predicted breast cancer incorrectly.
- **False Negative,** there are 84 patients who they are predicted recover incorrectly.

While the classification performance of breast cancer dataset using Naïve Bayes as follows in Table 9:

Table 9. Classification performance of breast cancer dataset using Naïve Bayes

Accuracy	Precision	Recall	F-measure
88.61%	92.76%	86.21%	89.36%

The information of Table 9 above is **precision** of each sample which calculate between patients who are attacked breast cancer correctly and patients who are predicted breast cancer incorrectly that reach 92.76%.

5.4.2 Decision Tree

Decision tree is an inductive learning task which is a predictive model based on branching series of Boolean tests. In chapter 4.4.2 was explained the conceptual of Decision Tree. Our research applied C4.5 to make decision tree model which is shown in Figure 16.

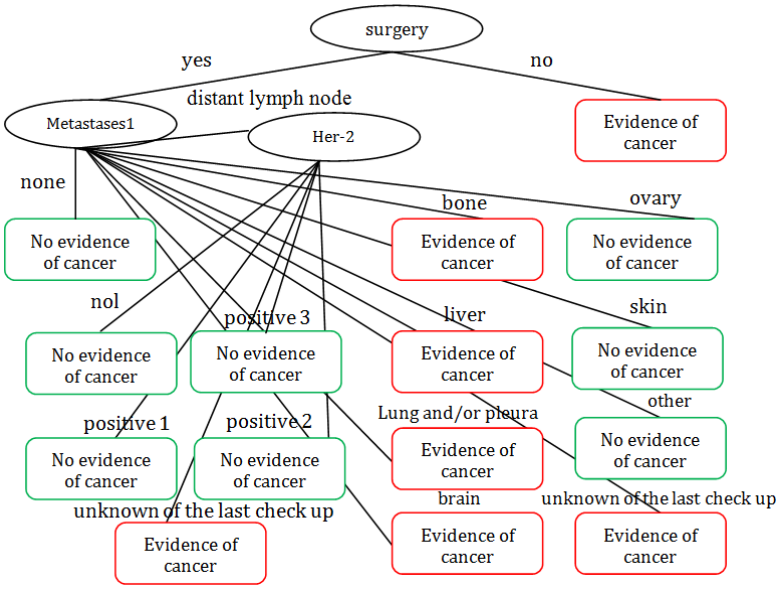

Figure 16. Decision tree model of breast cancer dataset

The Figure 16 above can be explained in 'IF-THEN' rules as follows in Table 10:

Table 10. IF-THEN rules of breast cancer dataset

RULE 1	if (surgery == 'yes' and metastase1 =='None') ==>**No evidence of cancer**
RULE 2	if (surgery == 'yes' and metastases1 == 'Distant lymph node' and HER-2 == 'nol') ==>**No evidence of cancer**
RULE 3	if (surgery == 'yes' and metastases1 == 'Distant lymph node' and HER-2 == 'positive1') ==>**No evidence of cancer**
RULE 4	if (surgery == 'yes' and metastases1 == 'Distant lymph node' and HER-2 == 'positive2') ==>**No evidence of cancer**
RULE 5	if (surgery == 'yes' and metastases1 == 'Distant lymph node' and HER-2 == 'positive3') ==>**No evidence of cancer**
RULE 6	if (surgery == 'yes' and metastases1 == 'Distant lymph node' and HER-2 == 'Unknown of the last check up') ==>**Evidence of cancer**
RULE 7	if (surgery == 'yes' and metastases1 == 'Bone') ==>**Evidence of cancer**
RULE 8	if (surgery == 'yes' and metastases1 == 'Liver') ==>**Evidence of cancer**
RULE 9	if (surgery == 'yes' and metastases1 == 'Lung and/or Pleura') ==>**Evidence of cancer**
RULE 10	if (surgery == 'yes' and metastases1 == 'Brain') ==>**Evidence of cancer**
RULE 11	if (surgery == 'yes' and metastases1 == 'Ovary') ==>**No evidence of cancer**
RULE 12	if (surgery == 'yes' and metastases1 == 'Skin') ==>**No evidence of cancer**
RULE 13	if (surgery == 'yes' and metastases1 == 'Other') ==>**No evidence of cancer**
RULE 14	if (surgery == 'yes' and metastases1 == 'Unknown of the last check up') ==>**Evidence of cancer**
RULE 15	if (surgery== 'no') ==>**Evidence of cancer**

The performance metric begin from TP, TN, FP, and FN of breast cancer dataset using Leave-one-out cross validation and Decision Tree as follows in Table 11:

Table 11.TP, TN, FP, and FN results of breast cancer dataset using Decision Tree

True Positive	True Negative	False Positive	False Negative
557	446	42	52

The information of explanation in Table 11 above is:

- **True Positive**, there are 557 patients who they are predicted breast cancer correctly.
- **True Negative,** there are 446 patients who they are predicted recover correctly.
- **False Positive,** there are 42 patients who they are predicted breast cancer incorrectly.
- **False Negative,** there are 52 patients who they are predicted recover incorrectly.

While the classification performance of breast cancer dataset using Decision Tree as follows in Table 12:

Table 12. Classification performance of breast cancer dataset using Decision Tree

Accuracy	Precision	Recall	F-measure
91.43%	92.99%	91.46%	92.22%

The information of Table 12above is **precision** of each sample which calculate between patients who are attacked breast cancer correctly and patients who are predicted breast cancer incorrectly that reach 92.99%.

5.5. The Interesting Facts of Breast Cancer Dataset and ModelEvalution

Breast cancer dataset of our research is combination between first diagnosis and subsequent diagnosis (could be last diagnosis) which it was taken from Oncology Hospital. The uniqueness of its dataset also represented the factors of patient that were attacked breast cancer and they were given several treatments, and then they were grouped by two classes consist of no evidence of cancer (recover) and evidence of cancer (not recover yet). Figure 17 below illustrated the procedure of breast cancer registration records in Oncology Hospital.

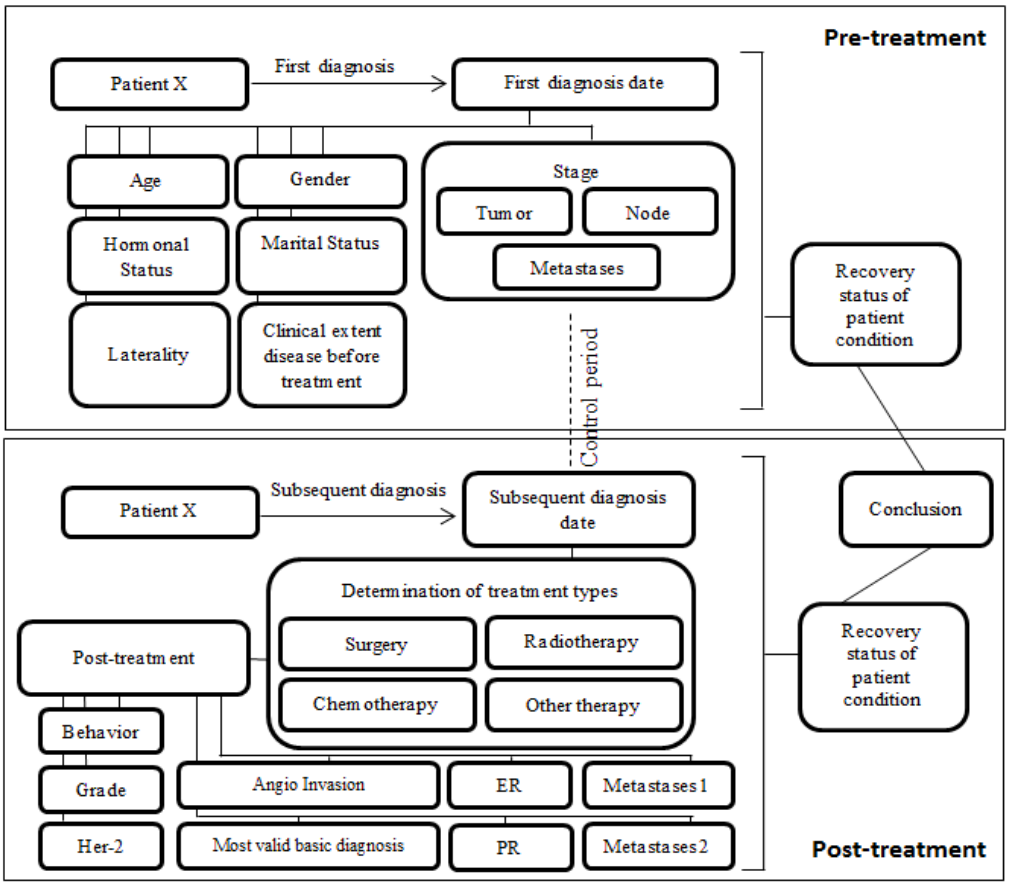

Figure 17. The procedure of breast cancer registration records in Oncology Hospital.

Figure 17 above illustrated how to create a breast cancer registration record of individual patient. Each data sample is taken from combination between pre-treatment (first diagnosis) and post-treatment (subsequent diagnosis or could be last diagnosis). Table 13 below is shown the factors that are recorded on each of pre-treatment and post-treatment process.

Table 13. The factors of pre-treatment and post-treatment process

Pre-treatment	Post-treatment
1. First diagnosis date 2. Gender 3. Marital status 4. Age 5. Stadium 6. Clinical extent disease before treatment 7. Laterality 8. Hormonal status	1. Behavior 2. Most valid basic diagnosis 3. Grade 4. Angio Invasion 5. ER 6. PR 7. HER-2 8. Surgery 9. Radiotherapy 10. Chemotherapy 11. Other therapy 12. Metastases 1 13. Metastases 2 14. Subsequent or last diagnosis date

Pre-treatment, Patients are first diagnosed their stage and other factor (8 factor records), and then the doctor will evaluate the patient control period for determining the next step to give:

- Medicines and outpatient, or
- Treatments.

Post-treatment, Patients will be given alternative breast cancer treatments. There are four types of treatment which they are offered in Oncology Hospital such as Surgery, Radiotherapy, Chemotherapy and Other Therapy. **Surgery** is the surgical removal of the cancer cells cut out part of the normal tissue, surgery is a local therapy.In our dataset, especially for surgery feature only is consist of two values ('yes' or 'no'), which it give information that 'yes' is patient who they using surgery in Oncology Hospital (our case study), while 'no' is patient who they using surgery in other hospital (not surgery in Oncology Hospital, e.g. general hospital, but theyusing chemotherapy, radiotherapy, and immunotherapy in Oncology Hospital), which Figure 18 shown the number of breast cancer patients in the Oncology Hospital of our research treated using surgery, (1) the patients who used surgery, there were 446 recover (free disease) of breast cancer or called no evidence of cancer patients, while 70 patients were not recovered or called evidence of cancer. In addition, (2) the patients who not used surgery, there were 42 no evidence of cancer patients, while 539 patients were evidence of cancer.

The second local therapy is radiotherapy. **Radiotherapy** is often given after breast conserving surgery to help lower the risk of a recurrence. Figure 19 shown the number of breast cancer patients in the Oncology Hospital treated using radiotherapy, (1) the patients who used radiotherapy, there were 138 no evidence of cancer patients, while 76 patients were evidence of cancer. In addition, (2) the patients who not used radiotherapy, there were

350 no evidence of cancerpatients, while 533 patients were evidence of cancer.

In addition to local therapy, there is also systemic therapy, the first is chemotherapy. **Chemotherapy** is the treatment of the entire body, it same with radiotherapy have strong side effects. Chemotherapy is often given after breast conserving surgery to help reduce the risk of recurrence. Figure 20 shown the number of breast cancer patients in the oncology hospital treated using chemotherapy, (1) the patientswho used chemotherapy, there are 352 no evidence of cancerpatients, while 169 patients were evidence of cancer. In addition, (2) patients who not used chemotherapy, there were 136 no evidence of cancer patients, while 440 patients were evidence of cancer.

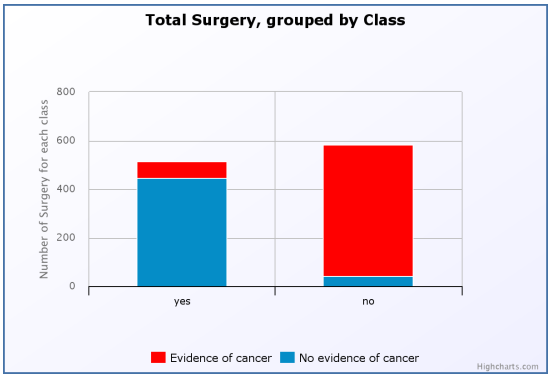

Figure 18. The number of breast cancer patients treated by surgery

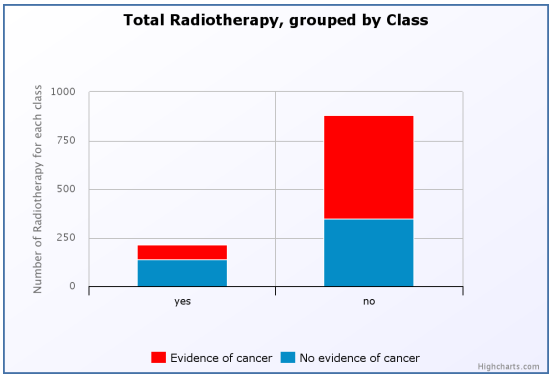

Figure 19. The number of breast cancer patients treated by radiotherapy

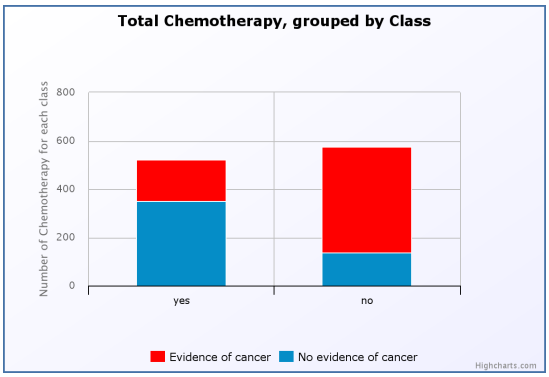

Figure 20. The number of breast cancer patients treated by chemotherapy

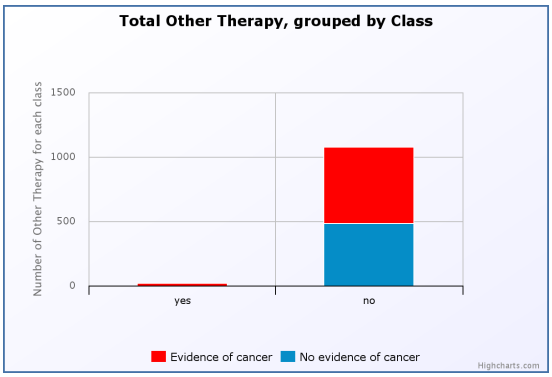

Figure 21. The number of breast cancer patients treated by other therapy

The second systemic therapy is immunotherapy and hormone therapy which both included as other therapy on data. **Immunotherapy** is a treatment to kill cancer cells, anti-recurrence and metastasis, reconstruct the immune system, while hormone therapy is most often used after surgery to reduce the risk of the cancer coming back. Figure 21 shown the number of breast cancer patients in the oncology hospital treated using other therapy, (1) the patientswho used other therapy,there are 3 no evidence of cancerpatients, while 15 patients were evidence of cancer. In addition, (2)

patients who not used other therapy, there were 485 no evidence of cancer patients, while 594 patients were evidence of cancer.

The Figure 18, 19, 20, and 21 described that the number of patients who used treatment were still low, even patients who were attacked by breast cancer that higher than patients who recover of breast cancer. The comparison between patients who are no evidence of cancer (recover) and evidence of cancer (not recover yet) that is 488:609.

After being given the treatment, and then the doctor will evaluate to the condition progress of patients, especially focus on factors:

- Metastases 1 (Spread of cancer in the area 1),
- Metastases 2 (Spread of cancer in the area 2), and
- If patients are recover (no evidence of cancer), then the treatment is considered successful. In other condition, if patients are not recover yet (evidence of cancer), then the doctor determines the next steps to the patients.

In model evaluation, supervised learning method has different characteristic model which it was generated by classification algorithm [4]. In conceptually, the medical dataset is useful as data source which the data mining expert will find the suitable classification method for several disease or health case studies, especially in decision making, diagnosis, prognosis, and finding pattern of disease.In our experiment result, the classification for determining patient who recover and not recover of Naïve Bayes and Decision Tree that has the different model. Naïve Bayes classification is tends to calculate the probabilistic of each feature value on each class. Therefore, the feature value which it has big portion in a class, it will be given opportunity for sample to be included in its class.On the other hand, Decision Tree is more representative than Naïve Bayes, because the model is easy to understand by human. Decision tree represent the tree shape which it is consist of node, trunk, branch, and leaf. The decision making conclusion of decision tree will generate “IF-THEN” rules.Based on our experiment used Naïve Bayes and Decision Tree, the precision is still in 92.76% and 92.99% respectively. However, it needs to apply the several classification methods for solving the classification model of breast cancer dataset in the further research.

6. CONCLUSION

We have successfully applied Information Gain, Fisher’s Discriminant Ratio and Chi-square as feature selection algorithm to breast cancer data. All three algorithms selected 14 most important features from 21 original features. Feature selection can be used to build powerful learning models. Hierarchical K-means clustering can help to determine the ideal number of features to be removed with error ratio and variance parameters. The three

feature selection methods given similar results to remove 7 features that have a low contribution in data are gender, marital status, hormonalstatus, other therapy, laterality, age and behavior. The accuracy results used 14 most important attributes given 81.68%, while the error ratio decreased from 44.48% to 18.32%. The 14 attributes were selected can useful to determinant factor of breast cancer patients on medical oncology. The comparison of classification algorithms between Naïve Bayes and Decision Tree were given precision reach 92.76% and 92.99% respectively by leave-one-out cross validation. Our research can discover the features that make a patient can be recovered, therefore the feature selection prediction result of our research can be recommended to another patient. But, we applied classification method still has an average precision between 92.76%-92.99%, we need further research to improve the precision until 7%. We also consult to doctor oncologist or medical experts to know the useful of our research can help their work. We got suggestions for improvement as follows:

- The feature selection method is helpful to analysis the features which have the high and low contribution that can impact to patient recovery factor.
- The classification method is a method which it tries to build decision-making, and to support the doctor work. But, the precision of prediction result must be improved, because it is related with human-life.
- The analysis of breast cancer is also can find through the relationship among individual features using specific data mining algorithm.

Furthermore, information based on our data research, the breast cancer patients in Indonesia especially in East Java must be improved the treatment routinely in the hospital to early recover of breast cancer.

REFERENCES

[1] Abdelghani B., Erhan G., **Predicting Breast Cancer Survivability using Data Mining Technique**s, *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining*, 2006.

[2] Albar ZA, Tjindarbumi D, Ramli M, Lukitto P, Reksoprawiro S,Handojo D,Darwis I, Suardi DR, Achmad D, **Protokol Peraboi 2003**, *PerhimpunanAhliBedah Onkologi Indonesia*, 2004.

[3] American Cancer Society 2011-2012, **Breast Cancer Survival Rates by Stage*:Breast Cancer Guidelines***, *American Cancer Society Breast Cancer Facts&Figures,* 2012.

[4] AndriPermanaWicaksono, TessyBadriyah, AchmadBasuki, **Comparison of The Data-Mining Methods in Predicting The Risk Level of Diabetes**, *EMITTER International Journal of Engineering Technology*, Vol. 4, No.1, pp. 164-178, 2016.

[5] Bustami, **Penerapan Algoritma Naïve Bayes untuk Mengklasifikasi Data Nasabah Asuransi,** *TECHSI: JurnalPenelitianTeknikInformatika*, Vol.8, No.1, pp.128-146, 2014.
[6] Charu C. Aggarwal, **Data Classification: Algorithms and Applications,** *CRC Press*, pp. 1-667, 2014.
[7] Cheng T.Y.,, Cheng M. C., Bor W. C., **Prediction of Survival in Patients with Breast Cancer using Three Artificial Intelligence Techniques**, *Journal of Theoretical and Apllied Information technology*, Vol.60, No.1, pp. 179-183, 2014.
[8] Cornain S, Mangunkusumo R, Nasar IM, Pribartono J, **Ten Most Frequent Cancers in lndonesia :Pathology based Cancer Registry Data of l988-1989**, *In:Cancer Registry in Indonesia, National Cancer Registry Center, JakartaCoordinating Board*, 1990.
[9] EkoPrasetyo, **Data Mining-Mengolah Data Menjadi Informasi Menggunakan Matlab**, *Andi Offset*, Ed.1, pp. 28-30, 2014.
[10] Hadi L. A., Maryam A., Masoud R., Farahnaz S., **Prediction of Breast Cancer Survival Through Knowledge Discovery in Databases**, *Global Journal of Health Science*, Vol.7, No.4, pp.392-398, 2015.
[11] International Agency for Research on Cancer, **Latest World Cancer Statistics Global Cancer Burden Rises to 14.1 Million New Cases in 2012: MarkedIncrease in Breast Cancers Must be Addressed**, *IARC Press ReleaseN° 223* , 2013.
[12] Irawan C, Hukom R, Prayogo N, **Factors Associated with Bone Metastasis in Breast Cancer: A Preliminary Study in An Indonesian Population**, *Acta MedIndones-Indones J Intern Med,* Vol.40, No.4, pp.178-180, 2008.
[13] IwanSyarif, **Feature Selection of Network Intrusion Data using Genetic Algorithm and Particle Swarm Optimization**, *EMITTER International Journal of Engineering Technology*, Vol. 4, No.2, pp. 277-290, 2016.
[14] Jaree T.,Guandong X., Yanchun Z., Fuchun H., **Breast Cancer Survivability via Ada Boost Algorithms**, *In: Health data and knowledge management: proceedings of the Second Australasian Workshop on Health Data and Knowledge Management (HDKM), Wollongong, NSW, Australia,* Vol. 80, pp.55-64, 2008.
[15] Jemal A, Clegg LX, Ward E, Ries LA, Wu X, Jamison PM, Wingo PA, HoweHL, Anderson RN, Edwards BK, **Annual Report to The Nation on The Status ofCancer, 1975-2001, with A Special Feature Regarding Survival**, *Cancer*, Vol.101, No.1, pp.3-27, 2004.
[16] Kohei Arai, Ali Ridho Barakbah, **Hierarchical K-means: an Algorithm forCentroids Initialization for K-means**, *Reports of the Faculty of Science andEngineering, Saga University, Japan*, Vol.36, No.1, 2007.
[17] Luis Carlos Molina, LluísBelanche, ÀngelaNebot, **Feature Selection Algorithms: A Survey and Experimental Evaluation**, *IEEE*

International Conference on Data Mining, Maebashi City, Japan, pp. 1 – 19, 2002.

[18] Mc Cready D, Holloway C, Shelley W, Down N, Robinson P, Sinclair S, MirskyD, **Surgical Management of Early Stage Invasive Breast Cancer: A Practice Guideline**, *Can J Surg*, Vol.48, No.3, pp.185-194, 2005.

[19] Ministry of Health Republic of Indonesia**, Indonesia Health profile 1993**, *Jakarta:DepartemenKesehatan RI,*1993.

[20] National Breast and Ovarian Cancer Centre, **Breast Cancer Risk Factors: A Review of The Evidence**, National Breast and Ovarian Cancer Centre, SurryHills, NSW,*Resources for Health Progessionals,* 2009.

[21] Ng CH, Pathy NB, Taib NA, Teh YC, Mun KS, Amiruddin A, Evlina S,Rhodes A, Yip CH, **Comparison of Breast Cancer in Indonesia and Malaysia**—aClinico-Pathological Study between Dharmais Cancer Centre Jakarta andUniversity Malaya Medical Centre, Kuala Lumpur, *Asian Pac J Cancer Prev*, Vol.12, No.11, pp.2943-2946, 2011.

[22] Prihartono J, Mangunkusumo R, Partoatmodjo P, **Establishing Pathology based Cancer Registry: Indonesian Experience**. In: Sasaki R, Aoki K, editors.Epidemiology and Prevention of Cancer. Proceedings of Monbusho (Ministryof Education, Science & Culture) International Symposium on CoparativeStudy of Etiology & Prevention of Cancer, Nagoya, 1989. *Nagoya: TheUniversity of Nagoya Press*, pp. 211-16, 1990.

[23] R. K. Kavitha, Dorai R, **Predicting Breast Cancer Survivability using Naïve Bayes Classifier and C4.5 Algorithm**, *Elysium Journal*, Vol.1, No.1, pp.61-63, 2014.

[24] Sergio Verdu, Fellow, IEEE, **Fifty Years of Shannon Theory**, *IEEE Transactions on Information Theory*, Vol.44, No.6, pp.2057-2078, 1998.

[25] Tjindarbumi, **Diagnosis dan Pencegahan Kanker Payudara,** *KursusSingkatDeteksiDinidanPencegahanKanker*, FKUI-POI, Jakarta, 6-8 November, 1995.

[26] Tresna, MaulanaFahrudin, IwanSyarif, Ali RidhoBarakbah, **Ant Colony Algorithm for Feature Selection on Microarray Datasets**,*The Eighteenth International Electronics Symposium (IES)-IEEE co-sponsored conference*, Bali, Indonesia, 2016.

[27] Tresna Maulana Fahrudin, Iwan Syarif, Ali Ridho Barakbah, **The Determinant Factor of Breast Cancer on Medical Oncology using Feature Selection Based Clustering**, *The Fifth International Conference on Knowledge Creation and Intelligent Computing (KCIC) 2016-IEEE co-sponsored conference*, Manado, Indonesia, 2016.

[28] Wakai K, Dillon DS, Ohno Y, Prihartono J, Budiningsih S, Ramli M, Darwis I,Tjindarbumi D, Tjahjadi G, Soetrisno E, Roostini ES, Sakamoto G, Herman S,Cornain S, **Fat Intake and Breast Cancer Risk in An Area Where Fat Intake isLow: A Case-Control Study in Indonesia**, Int J Epidemiol, Vol.29, No.1, pp.20-28, 2000.

[29] Zulaiha Ali Othman, Azuraliza Abu Bakar, Abdul RazakHamdan, Khairuddin Omar, Nor LiyanaMohd Shuib, **Agent Based Preprocessing**, *International Conference on Intelligent and Advanced Systems*, KL Convention Centre, pp. 219 – 223, 2007.