# Breast Cancer Classification with Weighted Decision Tree Approach

[1] Khin Thuzar Win, [2] Aung Nway Oo
[1] University of Computer Studies Hinthada, Myanmar
khinthuzarwin87@gmail.com
[2] University of Information Technology, Myanmar
aungnwayoo78@gmail.com

***Abstract*— Classification can be used as in the form of data analysis that can be used to extract models describing the important data classes. Classification is the task to identify the class labels for instances based on a set of features (attributes). This paper will present the traditional decision tree and weighted decision tree algorithms. In this study, C4.5 and CART decision tree algorithms are used to predict the breast cancer. Naïve Bayesian theorem was used to calculate the weight value to set the appropriate weights to decision tree model. The research work focuses the comparative analysis of weighted decision tree algorithms and traditional decision tree algorithms by using Breast Cancer datasets.**

***Keywords*** **- Classification, Decision Tree, Naïve Bayesian, Weighted Decision Tree**

## I. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amount of data. Data mining is the part of the knowledge discovery in databases. In today's computer-driven world, these databases contain massive quantities of information. The accessibility and abundance of this information makes data mining a matter of considerable important and necessity. Most data mining techniques are based on inductive learning, where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of inductive approach is that the trained model is applicable to future, unseen examples.

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms and machine learning methods (algorithms that improve their performance automatically through experience, such as neural network or decision tree). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. Data mining can be performed on data represented in quantitative, textual or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association, classification, clustering and forecasting.

Classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends whose class label is unknown. Classification can be used for making intelligent decisions. Many classification methods have been proposed by researches and are important for research and practical application in a variety of fields: including pattern recognition and artificial intelligence, statistics, vision analysis, medicine and so on.

Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Classification has been successfully applied to wide range of application areas, medical diagnosis, weather prediction, credit approval customer segmentation, fraud detection among the different proposals. Classification is clearly useful in many decision problems, where for a given data item a decision is to be made (which depend on the class to which the data item belongs).

Decision trees are commonly used in classification systems because they are easy to interpret, accurate, and fast. Decision trees are attractive because they show clearly how to reach a decision and they are easy to construct. There have been many decision tree algorithms like ID3 [1], C4.5 [2], CART [3] etc.

Classification can be used as in the form of data analysis that can be used to extract models describing important data classes. In this study, decision tree algorithms: C4.5, CART and weighted decision tree algorithms are implemented and an experiment is performed to compare their results obtained from both training and testing phases. The rest of the paper is organized as follows. Section 2 reviews the related work and section 3 presents the overview C4.5 algorithm. CART algorithm was described in section 4. Naïve Bayes theorem and weighted C4.5 algorithm were described in section 5 and 6. Weighted CART algorithm was described in section 7.Overview of the system flow was illustrated in section 8. The data set description and experimental results are presented in section 9 and 10. Finally, conclusion of this study was provided in section 11.

## II. RELATED WORK

There are many research works that relate to classification with weighted decision trees approach. [10]. Diagnosis of breast cancer with decision tree and artificial neural network was proposed in [11]. Puneet and group proposed breast cancer classification with decision tree model and SVM. Yamuna and Venkatesan [12] proposed the kidney transplant survival rate prediction with decision trees and comparative analysis is performed by using C4.5 and CART decision tree algorithms. E. Venkatesan* and T. Velmurugan [13] researched the performance analysis of decision trees algorithms for

breast cancer classification. They used the j48, AD tree and CART algorithms and then compared and evaluate the results of different classifiers. In the paper of Hyontai Sug [14] Comparison of Decision Tree Algorithms for Medical Data Sets were performed using the C4.5 and CART. To evaluate the algorithm, used the 17 medical datasets and 10 fold cross validation was performed and compare the accuracy of algorithms. According to literature survey C4.5 has been used in some wide range of areas [15] like financial areas [16] and engineering areas [17], but CART has been favored mostly in medicine domain, because most researchers in medicine domain reported good performance of CART in their data mining tasks [18].

In this paper, comparative studies of weighted decision trees and traditional decision tree are made to predict the breast cancer dataset.

## III. C4.5 ALGORITHM

The C4.5 algorithm is the modified version of ID3 algorithm and which choose splitting attributes from a dataset with the highest information gain.

Let $p_i$ be the probability that an arbitrary tuple in D belongs to class $C_i$, estimated by $|C_{i,\ D}|/|D|$. Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (1)$$

where $p_i$ is the probability that an arbitrary tuple in $D$ belongs to class $C_i$ and is estimated by $p_i = |C_i,D|/|D|$, where $|C_iD|$ is total tuple for $C_i$, and $|D|$ is total tuple.

Information needed (after using attribute A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \qquad (2)$$

where $|D_j|$ is total tuples in D that have outcome $a_j$ of A, and $|D|$ is total tuple. Information gained by branching on attribute A.

$$Gain(A) = Info(D) - Info_A(D) \qquad (3)$$

In other words, *Gain(A)* is the expected reduction in entropy caused by knowing the value of the attribute A. Information gain measure is biased towards attributes with a large number of values. C4.5 (a successor of ID3) uses gain ratio to overcome the problem (normalization to information gain)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \qquad (4)$$

SplitInfo(A) is the information due to the split of C on the basis of the value of the categorical attribute A. The attribute that yields the largest Gain Ratio is chosen for the decision node. The attribute with the maximum gain ratio is selected as the splitting attribute;

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \qquad (5)$$

The attribute with the highest information Gain Ratio is chosen as the test attribute at each node in the tree. Such a measure or a measure is the goodness of split. The attribute with the highest information gain ratio is chosen as the attribute for the current root node.

## IV. CART ALGORITHM

CART is capable of handling both numerical and categorical variables. Gini index measures how well a given attribute separates training samples into targeted class. Here binary splitting of attributes takes place. It is most widely used statistical procedure. It provides a hierarchy of binary decision [9].The Gini index is used in CART. If a data set $D$ contains examples from $m$ classes, gini index, $gini(D)$ is defined as

$$gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \qquad (6)$$

where $p_i$ is the relative frequency of class $i$ in $D$, *where* $p_i = |C_iD| / |D|$, $|C_iD|$ = total tuple for $C_i$ and $|D|$ = total tuple. The sum is computed over $m$ classes.

The Gini index considers a binary split for each attribute. Let's first consider the case where attribute $A$ is a discrete-valued attribute having $v$ distinct values, $\{a_1, a_2, \ldots, a_v\}$, occurring in $D$. Each subset, *SubA*, can be considered as a binary test for attribute $A$ of the form "$A$ 2 *SubA*?". Given a tuple, this test is satisfied if the value of $A$ for the tuple is among the values listed in *SubA*. If $A$ has $v$ possible values, then there are $2v$ possible subsets. Therefore, there are $2v$-2 possible ways to form two partitions of the data, $D$, based on a binary split on $A$.

If a data set $D$ is split on A into two subsets $D_1$ and $D_2$, the *gini* index $gini(D)$ is defined as

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2) \qquad (7)$$

where $|D_1|$ = the set of tuples in D satisfying A ≤ split-point, $|D_2|$ = the set of tuples in D satisfying A > split-point and $|D|$ = total tuple. Reduction in Impurity:

$$\Delta gini(A) = gini(D) - gini_A(D) \qquad (8)$$

The attribute provides the smallest $gini_{split}(D)$ (or the largest reduction in impurity) is chosen to split the node (*need to enumerate all the possible splitting points for each attribute*) [5].

Decision trees are formed by a collection of rules based on variables in the modeling data set [8]:

- Rules based on variables' values are selected to get the best split to differentiate observations based on the dependent variable
- Once a rule is selected and splits a node into two, the same process is applied to each "child" node (i.e. it is a recursive procedure)
- Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. (Alternatively, the data are split as much as possible and then the tree is later pruned.)

Each branch of the tree ends in a terminal node. Each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

## V. NAÏVE BAYES THEOREM

Naïve Bayesian (NB) classifier is a simple probabilistic classifier based on probability model, which

can be trained very efficiently in a supervised learning [12][4]. The naïve Bayesian classifier, or simple Bayesian classifier [5], works as follows:

1. Let $D$ be a training set of tuples and their associated class labels. As usual, each tuple is represented by an $n$-dimensional attribute vector, $X = (x_1, x_2, \ldots, x_n)$, depicting $n$ measurements made on the tuple from $n$ attributes, respectively, $A_1, A_2, \ldots, A_n$.

2. Suppose that there are $m$ classes, $C_1, C_2, \ldots, C_m$. Given a tuple, $X$, the classifier will predict that $X$ belongs to the class having the highest posterior probability, conditioned on $X$. That is, the naïve Bayesian classifier predicts that tuple $X$ belongs to the class $C_i$ if and only if

$$P(C_i|X) > P(C_j|X) \quad for\ 1 \le j \le m, j \ne i. \qquad (9)$$

Thus we maximize $P(C_i|X)$. The class $C_i$ for which $P(C_i|X)$ is maximized is called the *maximum posteriori hypothesis*.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (10)$$

3. As $P(X)$ is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C2) = \ldots = P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$\equiv \arg\max P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \qquad (11)$$

We can easily estimate the probabilities $P(x_1|C_i)$, $P(x_2|Ci)$, … , $P(x_n|Ci)$ from the training tuples. Weight value for each attribute is calculated by equation 8 which is the maximum weight value.

## VI. WEIGHTED C4.5 ALGORITHM

Weighted decision tree learning algorithm was developed by assigning appropriate weights to training instances, which improve the classification accuracy. The weights of the training instances are calculated using naïve Bayesian theorem. Weight of each training instance is calculated with the maximum value of the class conditional probabilities. Weighted C4.5 algorithm calculated the information gain by using these weights and builds the decision tree model for classification. Given a training dataset, the weighted C4.5 algorithm initializes the weights of each training instance, *Wi* by highest posterior probability for that training instance. The algorithm uses the weight value calculated from Naïve Bayes probabilistic model to initialize the weights of each training instance.

Let $p_i$ be the probability that an arbitrary tuple in D belongs to class $C_i$. Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (12)$$

Where, $p_i$ is the probability that an arbitrary tuple in $D$ belongs to class $C_i$ and it is calculate as: $p_i = \sum Wi / \sum_{j=1}^{n} |Wj|$, where, $W_i$ is weight for Class $C_i$, $W_j$ is weight for tuple j.

Information needed (after using attribute A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \qquad (13)$$

where $|D_j|$ is total weight tuples in D that have outcome $a_j$ of A, and $|D|$ is total weight tuple. We are calculated Gain(A) ,$SplitInfo_A(D)$ and GainRatio to assign weight value. The attribute with the highest information Gain Ratio is chosen as the test attribute at each node in the tree. Such a measure or a measure is the goodness of split. The attribute with the highest information gain ratio is chosen as the attribute for the current root node.

## VII. WEIGHTED CART ALGORITHM

Weighted decision tree learning algorithm was developed by assigning appropriate weights to training instances, which improve the classification accuracy. The weights of the training instances are calculated using naïve Bayesian theorem. Weight of each training instance is calculated with the maximum value of the class conditional probabilities. Weighted CART algorithm calculated the gini index by using these weights and builds the decision tree model for classification. Given a training dataset, the weighted CART algorithm initializes the weights of each training instance, *Wi* by highest posterior probability for that training instance. The algorithm uses the weight value calculated from Naïve Bayes probabilistic model to initialize the weights of each training instance.

If a data set $D$ contains examples from $m$ classes, gini index, $gini(D)$ is defined as

$$gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \qquad (14)$$

where $p_i$ is the relative frequency of class $i$ in $D$, *where* , $p_i = \sum Wi / \sum_{j=1}^{n} |Wj|$, $W_i$ = weight for Class $C_i$ and $W_j$ = weight for tuple j. The sum is computed over $m$ classes.

The Gini index considers a binary split for each attribute. Let's first consider the case where attribute $A$ is a discrete-valued attribute having $v$ distinct values, $\{a_1, a_2, \ldots, a_v\}$, occurring in $D$. Each subset, *SubA*, can be considered as a binary test for attribute $A$ of the form "$A$ 2 *SubA*?". Given a tuple, this test is satisfied if the value of $A$ for the tuple is among the values listed in *SubA*. If $A$ has $v$ possible values, then there are $2v$ possible subsets. Therefore, there are $2v$-2 possible ways to form two partitions of the data, $D$, based on a binary split on $A$.

If a data set $D$ is split on A into two subsets $D_1$ and $D_2$, the *gini* index $gini(D)$ is defined as

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2) \qquad (15)$$

where $|D_1|$ = the total weight of tuples in D satisfying A ≤ split-point, $|D_2|$ = the total weight of tuples in D satisfying A > split-point and $|D|$ = total weight tuple.

Reduction in Impurity:

$$\Delta gini(A) = gini(D) - gini_A(D) \qquad (16)$$

Attribute provides the smallest $gini_{split}(D)$ (or the largest reduction in impurity) is chosen to split the node

(*need to enumerate all the possible splitting points for each attribute*).

## VIII. SYSTEM FLOW OF PROPOSED SYSTEM

The system flow for classification of breast cancer dataset with weighted decision tree algorithms were described in the following figure.

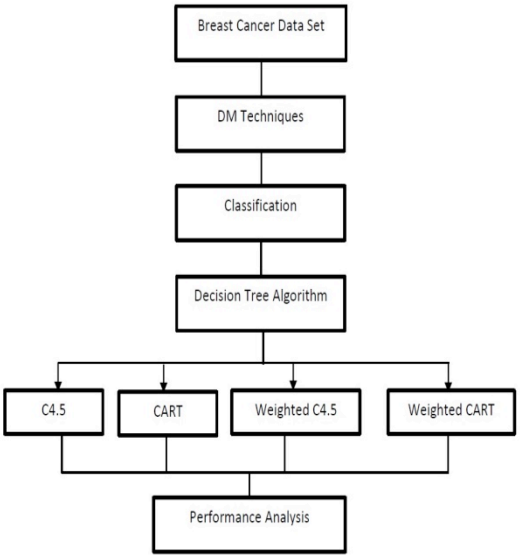


**Figure 1. Overview of the system flow**

## IX. DATASET DESCRIPTION

The breast cancer dataset contains 683 instances and 10 attributes. Each of the characteristics is assigned a value from 1 to 10 by the pathologist. The larger the value of attribute the greater the likelihood of malignancy.

The following table lists the attribute information of breast cancer dataset.

**Table 1. Dataset Description**

| ID | Attribute Name | Value |
|---|---|---|
| 1 | Clump Thickness | 1 – 10 |
| 2 | Uniformity of Cell Size | 1 – 10 |
| 3 | Uniformity of Cell Shape | 1 – 10 |
| 4 | Marginal Adhesion | 1 – 10 |
| 5 | Single Epithelial Cell Size | 1 – 10 |
| 6 | Bare Nuclei | 1 – 10 |
| 7 | Bland Chromatin | 1 – 10 |
| 8 | Normal Nucleoli | 1 – 10 |
| 9 | Mitoses | 1 – 10 |
| 10 | Class | Benign(2), or malignant(4) |

There are two types of classes in dataset, benign (It does not invade nearby tissue or spread to other parts of the body), or malignant (It is serious and likely to spread other parts of the body).

## X. EXPERIMENTAL RESULTS

The experimental results of classifiers are discussed in this section. The breast cancer dataset from UCI [7] is used for comparative analysis. For each classifier, 2/3 of the dataset is used for training and 1/3 of datasets is used for testing.

The following table compares the accuracy results of two classifiers.

| Algorithms | Recall | Precision | F-measure | Specificity | Accuracy (%) |
|---|---|---|---|---|---|
| C4.5 | 0.714 | 0.909 | 0.8 | 0.947 | 84.85 % |
| CART | 0.857 | 0.857 | 0.857 | 0.895 | 87.88 % |
| Weighted C4.5 | 1 | 1 | 1 | 1 | 100% |
| Weighted CART | 0.786 | 1 | 0.88 | 1 | 90.91 % |

The following figure visualizes the accuracy results of classifiers.

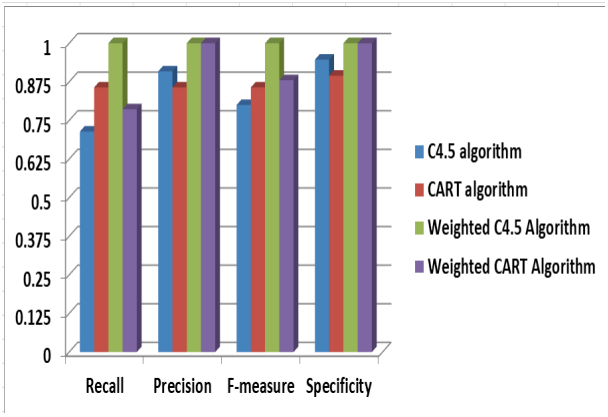


**Figure 2. Comparison of Recall, Precision, F-measure and Specificity**

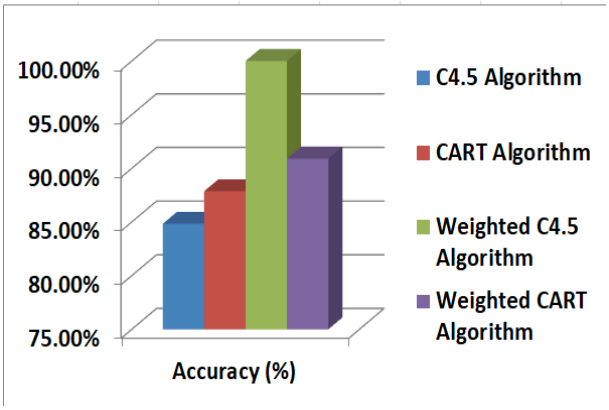


**Figure 3. Comparison of classification accuracy**

## XI. CONCLUSION

In this paper, the comparative analysis of the traditional decision tree and weighted decision tree algorithms on Breast Cancer classification problems. By comparing classification results, we confirm that weighted C4.5 algorithm is better than other classification algorithm for Breast Cancer classification dataset. The experimental results proved that the weighted C4.5 algorithm is more suitable for prediction of breast cancer dataset.

## ACKNOWLEDGMENT

I would to express my gratitude to University of Computer Studies Yangon for allow me to do this research work. Thanks to Dr. Aung Nway Oo for discussions that helped clarify our ideas and his support and encouragement. The author carried out this work while at Yangon Technological University and the Institute for the Study of Learning and Expertise.

## REFERENCES

[1]. J. R. Quinlan, "Induction of Decision Tree," Machine Learning Vol. 1, 1986, pp. 81-106.
[2]. J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, San Mateo,CA, 1993.
[3]. L. Breiman, J. H. Friedman, R. A. Olshen and C.J. Stone, "Classification and Regression Trees,"Statistics probability series, Wadsworth, Belmont, 1984.
[4]. Langely, P., Iba, W., Thomas, and K., "An analysis of Bayesian classifier," in Proceedings of the 10th national Conference on Artificial Intelligence (San Matro, CA: AAAI press), 1992, pp. 223-228.
[5]. Han, Jiawei and Kamber, Micheline "Data Mining Concepts and Techniques" 2nd ed., Morgan Kaufmann Publishers, San Franciso, CA, 2007 ISBN 1-55860-901-3.
[6]. P. Hamsagayathri, P. Sampath "DECISION TREE CLASSIFIERS FOR CLASSIFICATION OF BREAST CANCER", International Journal of Current Pharmaceutical Research, ISSN- 0975-7066, Vol 9, Issue 2, 2017.
[7]. UCI Machine Learning Repository: "Breast Cancer Wisconsin (Original) Data Set", Dr. WIlliam H. Wolberg (physician) University of Wisconsin Hospitals Madison, is consin, USA , Donor: Olvi Mangasarian (mangasarian '@' cs.wisc.edu) Received by David W. Aha (aha '@' cs.jhu.edu)
[8]. https://en.wikipedia.org/wiki/CART_algorithm
[9]. E.Venkatesan and T.Velmurugan, "Performance Analysis of Decision Tree Algorithms for Breast Cancer Classificaiton", Indian Journal of Science and Technology, Val 8(29), November 2015.
[10]. Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta" Diagnosis of Breast Cancer using Decision Tree Models and SVM", International Research Journal of Engineering and Technology(IRJET), Mar-2018
[11]. Autsuo Higa, "Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Network Algorithms", International Journal of Computer Applications Technology and ResearchVolume 7–Issue 01, 23-27, 2018
[12]. A Comparative Analysis of Decision Tree Methods to Predict Kidney Transplant Survival, International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Volume 5, No. 3, March-April 2014, Page No. 225-229.
[13]. E. Venkatesan* and T. Velmurugan , Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification, Indian Journal of Science and Technology, ISSN (Online) : 0974-5645, Vol 8(29), Page No. 1-8.
[14]. Hyontai Sug, Performance Comparison of Decision Tree Algorithms for Medical Data Sets, International Journal of Mathematics and Computers in Simulation, ISSN No: 1998-0159, Volume 8, 2014, Page No. 107-115.
[15]. S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, 2007, Page No. 249-268.
[16]. Z. Chang, "The application of C4.5 algorithm based on SMOTE in financial distress prediction model," in Proceedings of 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011, Page No. 5852-5855.
[17]. S. Gao, "The Analysis and Application of the C4.5 Algorithm in Decision Tree Technology," Advanced Materials Research, vol. 457-458, 2012, Page No.754-757.
[18]. R.J. Lewis, An Introduction to Classification and Regression Tree (CART) Analysis, Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California..
[19]. Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta" Diagnosis of Breast Cancer using Decision Tree Models and SVM", International Research Journal of Engineering and Technology(IRJET), Mar-2018
[20]. Autsuo Higa, "Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Network Algorithms", International Journal of Computer Applications Technology and ResearchVolume 7–Issue 01, 23-27, 2018
[21]. A Comparative Analysis of Decision Tree Methods to Predict Kidney Transplant Survival, International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Volume 5, No. 3, March-April 2014, Page No. 225-229.
[22]. E. Venkatesan* and T. Velmurugan , Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification, Indian Journal of Science and Technology, ISSN (Online) : 0974-5645, Vol 8(29), Page No. 1-8.
[23]. Hyontai Sug, Performance Comparison of Decision Tree Algorithms for Medical Data Sets, International Journal of Mathematics and Computers in Simulation, ISSN No: 1998-0159, Volume 8, 2014, Page No. 107-115.
[24]. S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, 2007, Page No. 249-268.
[25]. Z. Chang, "The application of C4.5 algorithm based on SMOTE in financial distress prediction model," in Proceedings of 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011, Page No. 5852-5855.
[26]. S. Gao, "The Analysis and Application of the C4.5 Algorithm in Decision Tree Technology," Advanced Materials Research, vol. 457-458, 2012, Page No.754-757.
[27]. R.J. Lewis, An Introduction to Classification and Regression Tree (CART) Analysis, Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California..