Review Article **Biotechnology**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

MEASURE OF INEQUALITY AND IMPROVED CLASSIFICATION ACCURACY USING PMTCT HIV DATA

S.SANTHOSH KUMAR

Department of Computer Science, Government College for Women, Kumbakonam, Tamil Nadu

ABSTRACT

HIV (Human Immunodeficiency Virus) is a lent virus that causes Acquired Immune Deficiency Syndrome (AIDS). It is an epidemic which causes a progressive failure of immune system in the human body. All countries are maintaining HIV data repositories to monitor, share the needful information. For detailed investigation and diagnosis processes, the HIV data need to be categorized based on its generic variability, sub types and mode of transmission. The transmission of HIV mother to her child is one of the major causes of HIV transmission. It happens during pregnancy, labour, delivery and breastfeeding is called mother-to-child transmission One of important goal is to prevent mother to child HIV transmission (PMTCT). It is a global commitment accelerated by UNICEF, UNO, WHO and other organizations. These initiatives started very recent and hence the least amount of data is available about PMTCT. Mostly collected (PMTCT) data are with fewer interventions. When it is used for various clinical implications, it gives poor results with least accuracy. This paper focuses to handle the inconsistent data related to accuracy rate.

KEYWORDS: **HIV, C4.5, Gini Index, accuracy, error rate.**

S.SANTHOSH KUMAR
Department of Computer Science, Government College for Women, Kumbakonam, Tamil Nadu

***Corresponding author**

INTRODUCTION

Time and availability are two important aspects of data. Searching of information and gaining knowledge are the paths to deserve it. The HIV virus that causes AIDS is one of the most serious health challenges. The World Health Organization (WHO) statistics taken in the year 2013 which reveals that there were approximately 35 million people are living with HIV/AIDS. Of these, 3.2 million (nearly 240,000) were children between 0-15 years of age. According to WHO 2.1 million individuals were newly infected by HIV. This includes over 240,000 children less than15 years. Most of these children live in sub-Saharan Africa and were acquired by their HIV-positive mothers during pregnancy, birth or breastfeeding. A UNAID [3] report reveals that 19 million out of 35 million people living with HIV without knowing that they are infected. The scaling-up of PMTCT interventions in various regions are generally slow. The Preliminary prevention initiatives are failed; caused the women who are usually infected by their spouses. The awareness for family planning among all women, including HIV-infected women, is not widely available. HIV testing for pregnant women is not universally offered and coverage for the region remains least in the world. Consequently, coverage of pregnant women receiving the most effective antiretroviral regimen for PMTCT in 2010 was less than 5%. This reveals that the least availability of PMTCT data gets more significance[1-3]

PROPOSED WORK

The PMTCT data sets are inconsistent nature with irrelevant attributes with mixed values. Each dataset contains thousands of records with multiple attributes. To categorize PMTCT data sets; a classification technique is proposed, hence C4.5 is a widely used classifier which handles both sequential and discrete values. Even though C5.0 is developed for business purpose, C4.5 gives cost effective result with normalized information gain. It uses statistical classification technique which splits the whole dataset into categories with a new observation (training set). Based on training set the generic relationships of remaining data instances are measured. To obtain better consistency and accuracy in PMTCT data sets selection of primary attributes, missing values and error rate.

C4.5 CLASSIFIER

C4.5 is an algorithm developed by Ross Quinlan that generates Decision Trees (DT), which can be used for classification problems. It improves the ID3 algorithm by dealing with both continuous and discrete attributes, missing values and pruning trees after construction. C4.5 builds decision trees from a set of training data using information entropy. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The C4.5 algorithm then recurs on the smaller sub lists. This algorithm has a few base cases[4-6]

C4.5 Algorithm for building decision trees

Take all attributes as input
Generate selection criteria (training set)
Find the entropy
(for each attribute with highest information gain)
Create decision tree based on information gain
until the selection criteria in met
Repeat the same process until all attributes normalized

C4.5 WITH MISSING VALUES

C4.5 handles missing values effectively. It deals missing values with imputation which means the retrieval of missing feature of data with the use of available data. Distribution-based imputation is done when the dataset is split into multiple instances, each with a different value for the missing feature. It is assigned a weight corresponding to the estimated probability for the particular missing value and the weights sums up to 1.The splitting criterion is based on the difference in entropy. The attribute with the highest normalized information gain is chosen to make the decision[7].

$$IG\,(T, a) = H\,(T) - \sum_{v} \frac{|\{x \in T \mid x_a = v\}|}{T} . H\,(\{x \in T \mid x_a = v\})$$

The training data T is a set classified as S=S_1, S_2, S_3 . . . . Sn samples. Each sample s_i consists of a p-dimensional vector $x_{1,i}$, $x_{2,i}$, ... $x_{p,i}$, where x_i represent attributes or features of the sample, as well as the class in which s_i falls.

(i) Note that the information gain and entropy helps to rank all attributes in to subsets. The selected subsets are closest to training set rather it is not closest to true results.
(ii) This phenomenon states that there must be invariant results that mean an amount of impurity or error in the classified attributes. The amount of error must be calculated to find inequality in data.

PMTCT DATA SET

The data sets are obtained from public health department in England a division of British government. The data set contains information about Children born in the United Kingdom to HIV-infected women (PMTCT). The dataset is categorized into infection status of child, time, period and region of birth. The dataset is partitioned in two tables. The **table 1** contains regional information of child birth to infected women from the year 2002 to 2013.

Table 1
Child Birth rate of HIV Infected women (PMTCT)

Country and region of birth	Number of births to HIV-infected women		
	2002 or earlier	2003-2013	Total
England			
North East	18	218	**236**
Cumbria and Lancashire	19	79	**98**
Yorkshire & the Humber	80	849	**929**
Greater Manchester	53	576	**629**
Cheshire and Merseyside	25	207	**232**
East Midlands	55	644	**699**
West Midlands	91	1,061	**1,152**
Anglia and Essex	44	519	**563**
South Midlands and Hertfordshire	83	701	**784**
London	2,063	5,693	**7,756**
Surrey, Sussex and Kent	95	524	**619**
Thames Valley	75	422	**497**
Wessex	33	230	**263**
Devon, Cornwall and Somerset	27	232	**259**
Avon, Gloucestershire and Wiltshire	14	119	**133**
Not Reported	78	6	**84**
England total	**2,853**	**12,080**	**14,933**
Wales	28	206	**234**
North Ireland	6	86	**92**
Scotland	266	401	**667**
Channel Islands and Isle of Man	8	8	**16**
UK total	**3,191**	**12,790**	**15,981**

The table 2 contains the information about infection status of children (up to age 16).

Table 2
Infection Status of children (PMTCT)

Country and region of birth	Infection status of children	
	Indeterminate	Uninfected
England		
North East	15	216
Cumbria and Lancashire	10	79
Yorkshire & the Humber	105	794
Greater Manchester	99	497
Cheshire and Merseyside	26	193
East Midlands	112	563
West Midlands	175	934
Anglia and Essex	131	420
South Midlands and Hertfordshire	142	611
London	886	6,347
Surrey, Sussex and Kent	97	486
Thames Valley	85	389
Wessex	46	210
Devon, Cornwall and Somerset	31	214
Avon, Gloucestershire and Wiltshire	51	72
Not Reported	2	16
England total	**2,013**	**12,041**
Wales	38	184
Northern Ireland	18	74
Scotland	51	567
Channel Islands and Isle of Man	0	13
UK total	**2,154**	**12,882**

LIMITATIONS OF C4.5 WITH PMTCT DATASET

The nature (tree) of C4.5 algorithm is related to biology. The inherent feature of biological data set contains sequential and interval data which makes suitable to use C4.5 classifier. The features of C4.5 are handles numeric attributes, deal sensibly with missing values and noisy data.

Table 3
Confusion matrix of Child Birth rate of HIV Infected women

Confusion matrix	North East	Cumbria and Lancashire	Yorkshire & the Humber	Greater Manchester	Cheshire and Merseyside	East Midlands	West Midlands	Anglia and Essex	South Midlands and Hertfordshire	London	Surrey, Sussex and Kent	Thames Valley	Wessex	Devon, Cornwall and Somerset	Avon, Gloucestershire and Wiltshire	Not Reported	Wales	Northern Ireland	Scotland	Channel Islands and Isle of Man	Sum
North East	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Cumbria and Lancashire	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Yorkshire & the Humber	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Greater Manchester	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Cheshire and Merseyside	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
E. Midlands	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
West Midlands	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Anglia and Essex	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
South Midlands and Hertfordshire	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
London	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Surrey, Sussex and Kent	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Thames Valley	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Wessex	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Devon, Cornwall & Somerset	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Avon, Gloucestershire and Wiltshire	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Not Reported	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Wales	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Nor. Ireland	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Scotland	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Islands and Isle of Man	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Sum	5	0	5	0	5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	

The technique used to handle missing values in C4.5 is to use of probability values for missing value rather assigning existing most common values of that attribute. This probability values are calculated from the observed frequencies in that instance. But in the case of given dataset, each values are collected from different regions and missing values cannot be regenerated with the use of existing values. When missing values are not determined then it is considered as error or empty. In the case of data set classification, the selection of high residual attribute leads poor quality and accuracy rate. The **Table 3** shows the confusion matrix with regional distribution of values. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The **Table 4** shows precision and recall is the fraction of relevant instances that are retrieved. The overall error rate is calculated as 0.8000.

Table 4
Attribute Performances

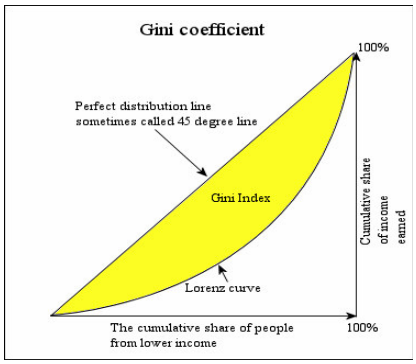
Value	Recall	1-Precision
North East	1.0000	0.8000
Cumbria and Lancashire	0.0000	1.0000
Yorkshire & the Humber	1.0000	0.8000
Greater Manchester	0.0000	1.0000
Cheshire and Merseyside	1.0000	0.8000
East Midlands	0.0000	1.0000
West Midlands	1.0000	0.8000
Anglia and Essex	0.0000	1.0000
South Midlands and Hertfordshire	0.0000	1.0000
London	0.0000	1.0000
Surrey, Sussex and Kent	0.0000	1.0000
Thames Valley	0.0000	1.0000
Wessex	0.0000	1.0000
Devon, Cornwall and Somerset	0.0000	1.0000
Avon, Gloucestershire and Wiltshire	0.0000	1.0000
Not Reported	0.0000	1.0000
Wales	0.0000	1.0000
Northern Ireland	0.0000	1.0000
Scotland	0.0000	1.0000
Channel Islands and Isle of Man	0.0000	1.0000

Note that both precision and recall of given dataset shows relevance of attributes based on an understanding and measure. The measure of inequality and error rate of each attribute is not accurately measured. The individual inequality measure of an attribute helps to identity the maximum strength of attribute selection (train set) and its weakness (error or missing). Therefore the inequality measure to be significant and selective measures must used.

MEASURE OF INEQUALITY

The Gini index is the most commonly used measure of inequality. It was developed by the Italian statistician and sociologist Corrado Gini. It measures the inequality among values of a frequency distribution.

Figure 1
Gini coefficient graph



A Gini coefficient is used to measure statistical dispersion. Its value of zero expresses perfect equality, where all values are the same and value of one expresses maximal inequality among values.

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

Where p (j | t) is the relative frequency of class j at node t. Hence the calculation of Gini index helps to predict and reduce the error rate. From the given attribute values, the inequality measure of an each attribute is determined. For that, the values of HIV infected women and infective statuses of child are compared to HIV infected.

Based on given values the missing values are calculated for each region of England, then the missing values are considered with actual values of infective status of children to obtain error rate of each attribute are shown in **Table 5.**

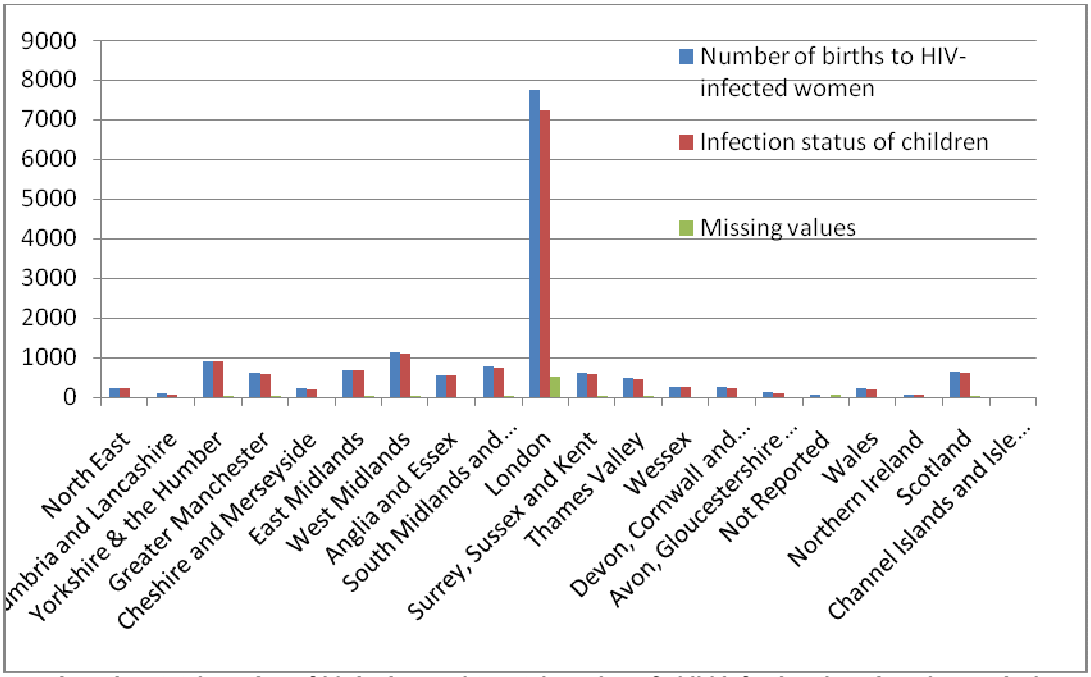
Table 5
Gini Index with missing values and error rate

Country and region of birth	Missing Values	Gini Index with Country Codes	Error Rate
England			
North East	5	$G_{NE} = 1-[(231/236)^2 + (5/236)^2]$	*0.042*
Cumbria and Lancashire	9	$G_{CL} = 1-[(79/98)^2 + (10/98)^2]$	0.340
Yorkshire & the Humber	26	$G_{YH} = 1-[(794/929)^2 + (105/929)^2]$	0.701
Greater Manchester	33	$G_{GM} = 1-[(497/629)^2 + (99/629)^2]$	0.351
Cheshire and Merseyside	13	$G_{CM=1-}[(193/232)^2 + (26/232)^2]$	0.704
East Midlands	23	$G_{EM} = 1-[(563/699)^2+(112/699)^2]$	0.350
West Midlands	43	$G_{WM} = 1-[(934/1152)^2 + (175/1152)^2]$	0.320
Anglia and Essex	12	$G_{AE} = 1-[(420/563)^2 + (131/563)^2]$	0.945
South Midlands and Hertfordshire	31	$G_{SMH=1-}[(611/784)^2 + (142/784)^2]$	0.360
London	523	$G_{L} = 1-[(6347/7756)^2 + (886/7756)^2]$	0.318
Surrey, Sussex and Kent	36	$G_{SSK} = 1-[(486/619)^2 + (97/619)^2]$	0.360
Thames Valley	23	$G_{TV} = 1-[(389/497)^2 + (85/497)^2]$	0.235
Wessex	0	$G_{W} = 1-[(210/259)^2 + (46/259)^2]$	0.820
Devon, Cornwall and Somerset	14	$G_{DCS} = 1-[(214/259)^2 + (31/259)^2]$	0.303
Avon, Gloucestershire and Wiltshire	10	$G_{AGW} = 1-[(72/133)^2 + (51/133)^2]$	0.560
Not Reported	66	$G_{NR} = 1-[(16/84)^2 + (2/84)^2]$	0.963
Wales	12	$G_{w} = 1-[(184/234)^2 + (38/234)^2]$	*0.356*
Northern Ireland	0	$G_{WL} = 1-[(74/92)^2 + (18/92)^2]$	*0.046*
Scotland	49	$G_S=NI-[(567/667)^2 + (51/667)^2]$	*0.272*
Channel Islands and Isle of Man	3	$G_{CIM} = 1-[(13/16)^2 + (0/16)^2]$	*0.340*

DISCUSSION

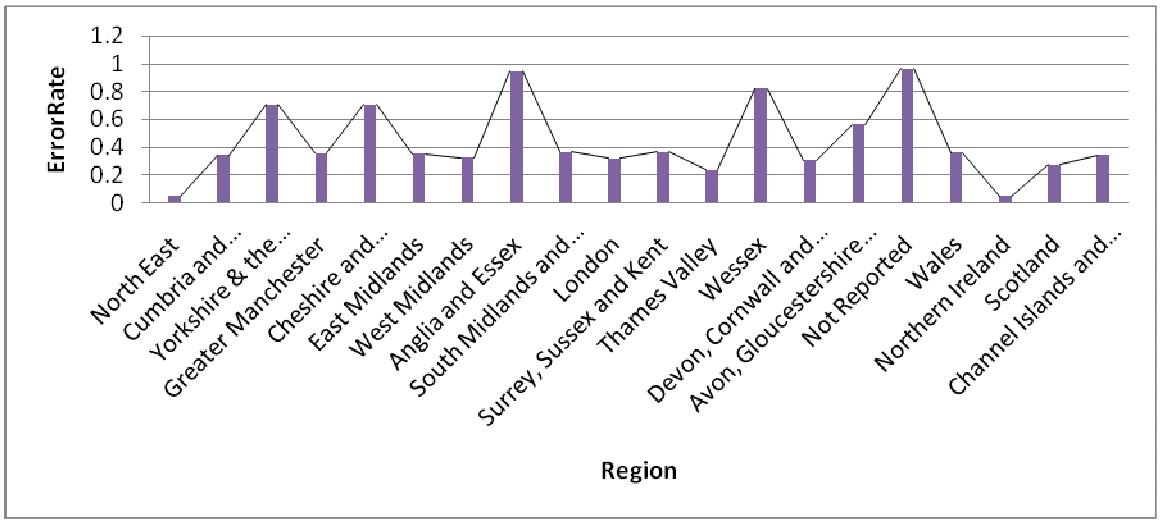
The proposed work is experimented with the use of existing classification and measuring techniques. For analysis, the HIV (PMTCT) data sets are used to measure inequality and error. The missing value ratio with respect to HIV infected women and infection status of children. The ratio of missing values with respect to child infection status depends on the total number of values available with missing value information.

Figure 2
Missing values of attributes



The figure 2 shows that when total number of births is equal to total number of child infection then there is no missing value therefore it is more accurate and considered as significant attribute. From Gini index, the highest missing values attribute is G_L but its volume is high similarly G_{NE} attribute achieves very less missing values. The figure 2 shows number births, infection status and missing values respectively. The attributes with highest volume with least error rate is suggested and preferable for significant attribute selection.

Figure 3
Error rate of attributes



The important point is the missing value of an attribute and its error rate is entirely different. The error rate is calculated purely based on the total number of births with child infection status. The value that remains not available in missing value and infection child status then it is considered to be an error. The error rate in shown in **Figure 3** which states that G_{NR} attribute achieved highest error rate 0.96, G_{AE} achieved 0.945 and G_W achieved 0.82 respectively. This result proves that the highest error rate attributes are not considered for primary attribute selection. The remaining attributes are having both minimum error rate and missing values so based on the relativity at the time of requirement other attributes are used.

CONCLUSION

Thus the work stated to handle the inconsistency in data sets and transformation and focused to study about inequality measure and error rate in PMTCT HIV data sets. The C4.5 classifier given better results, but it fails to achieve better performance in dealing with irrelevant attributes. The entropy measures using GINI index calculation given improved results with least error rate. The proposed techniques were also handled missing values in a data set for each attribute separately. The proposed technique enables to handle datasets with inconsistent attributes effectively. The proven results shows that the possibility that a minimum attribute value with least error rate attributes are not considered for train set. The highest value with least error rate attribute must be considered. The result gives redefined multi dimensional datasets with high consistency and high accuracy rate. The proposed approach can be used as a preliminary process for clinical data sets. This methodology can also be used to measure consistency and accuracy rate of data set with respect to dimensionality.

REFERENCES

1. Strachan M., Kwateng-Addo A., Hardee K., Subramaniam S., Judice N., Agarwal K. An Analysis Of Family Planning Content In Hiv/Aids, Vct, And Pmtct Policies In 16 Countries By January (2004). Accessed on 29 June 2015. http://www.policyproject.com/pubs/workingpapers/wps-09.pdf.
2. PillayY., Mametja D., Mbengashe T: Joint Review of HIV, TB and PMTCT Programmes in South Africa Main Report April (2014). Accessed on 2nd July 2015. http://www.hst.org.za/publications/joint-review-hiv-tb-and-pmtct-programmes-south-africa-april-2014.
3. Towards the elimination of mother-to-child transmission of HIV Conceptual framework for the Middle East and North Africa Region, World Health Organization (2012). Accessed on 3rd July. http://www.emro.who.int/images/stories/asd/documents/eMTCT_framework_-_EN_-_FINAL_WEB_-_26_Sep_2012.pdf.
4. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, (1993). Accessed on 22nd July. http://link.springer.com/article/10.1007%2FBF00993309#page-1.
5. Quinlan J.R. Improved use of continuous attributes in C4.5. J Artificial Intelligence Res, 4:77-90, (1996).
6. santhosh kumar .s., Ramaraj E. Analysis of Sequence Based Classifier Prediction for HIV Subtypes, Int J Engineering Tech, 2;1753-1758, (2012).
7. santhosh kumar .s., Ramaraj.E. Modified C4.5 Algorithm with Improved Information Entropy and Gain. Int J Engineering Res Tech, 3;1753-1758, (2013).
8. Frank A.F. The Gini Index and Measures of Inequality, The Mathematical Association Of America, December (2010). Accessed on 26th July. http://math.scu.edu/~ffarris/MonthlyFinal.pdf.
9. Dube A. Machine Learning Model for HIV1 and HIV2 enzyme secondary structure classification, J Comput Method Mol Design, 1 (2): 1-8, (2011).
10. Kotsiantis S.B., Supervised Machine Learning: A Review of Classification Techniques, Informatica, 31: 249-268, (2007).
11. Minghao P., Jong Burn L., Khalid.E.K., Keun H.R. Discovery of Significant classification rules from Incrementally Induced Decision Tree Ensemble for Diagnosis of Disease. Advanced Data Mining and Applications, 5678: 587-594 ,(2009).
12. Adhatrao K., Gaykar A., Dhawan A., Jha R., Honrao R. Predicting students' performance using Id3 and C4.5 classification algorithms, Int J Data Mining & Knowledge Management Process, 3(5); 39-52,(2013).