International Journal of Computational and Applied Mathematics. ISSN 1819-4966 Volume 12, Number 1 (2017)

ANALYSIS OF STUDENTS' BEHAVIOUR AND LEARNING USING CLASSIFICATION OF DATA MINING METHODS

K.Kiruthika[1], S. Sivakumar[2]

[1]*M.Phil, Department of Computer Science, C.P.A. College, Bodinayakanur, kirthi1790@gmail.com*

[2]*Associate Professor and Head, Department of Computer Science, C.P.A. College, Bodinayakanur, sivaku2002@yahoo.com*

ABSTRACT

Educational data mining concerns with developing the methods for discovering knowledge from data that come from educational environment. In this paper we used educational data mining to analyze learning behavior of modern students. In our case study, we collected students' data from a higher education institution in terms of the mentioned attributes. After preprocessing the data, we applied data mining techniques to discover classification, clustering and the algorithms such as AD Tree, J48 Tree and Random Tree. The hierarchical clusters were grouped with related similarities and analysis was experimented using WEKA tool. In each of these tasks, we extracted knowledge that describes students' learning behavior between traditional learning and virtual learning. The comparative performance of the above said algorithms are quantified with the values of stratified cross validation (SCV) are tabulated and illustrated in charts.

Key Words: *Educational Data Mining, E-Learning, Learning Management Systems, Class Technology.*

1. INTRODUCTION

Data mining in higher education is a new emerging field. Educational Data Mining (EDM) is concerned with developing the methods of exploring the unique types of data that come from educational settings. Through EDM the education has been benefited from a real E-Revolution as Virtual Learning Environment (VLE).The Internet also helped e-learning through the e-resources available for both teachers and students to share information. The goals of a teacher have changed from teaching facts into helping students to learn and how to find relevant information, how to access it, how to organize in different ways. Meanwhile, learning has moved towards more student-centre, problem-based, challenge-based or cooperative learning.

In traditional learning, the teacher is instructed to teach or facilitates all the sessions in classroom. In this the teacher usually talks more than the student. The teacher teaches the lesson according to the study program with the prescribed curriculum. In traditional learning the teachers take more effort than students, which had become unsuitable to today's curriculums. In Virtual Learning Environment, the relationship between the teacher and the learner is more effective through the beloved categories.

- **Learning achievement**
  Students in the VLE achieve better learning performance than in the traditional environment
- **Self-efficacy**
  Students in the VLE report higher levels of computer self-efficacy than there in the traditional environment
- **Satisfaction**
  Students in the VLE report higher levels of satisfaction than students in the traditional environment
- **Learning methodology**
  Students in the VLE report higher levels of learning climate than there in the traditional environment.

Literature Survey

Trilok Chand Sharma, Manoj Jain [2], the article titled as "WEKA Approach for Comparative Study of Classification Algorithm" discussed about classification of different decision tree algorithm as data mining techniques to process a dataset and identify the relevance of classification test data. Lukman, R., &Krajnc, M. (2012) [3], the article titled as "Exploring Non-traditional Learning Methods in Virtual and Real-world Environments". It represents an important step for the university and its further performance using non-traditional learning methods, due to the fact that most of the lectures carried out at the university are still done in a traditional way (lecturer-centred). In the virtual-class, more efforts should be directed towards decreasing frustrations by improving motivation and interactivity. Motivation could be strengthening by creating a sense of community and by building trust between students. Peter W. Stonebraker and James E. HazeltineNorth-eastern Illinois University, Chicago, Illinois, USA [4] titled as "Virtual learning effectiveness an examination of the process "This study defines, examines, and measures the effectiveness of a corporate virtual learning program. Initially, distinctions between traditional and virtual learning and university and corporate programs are defined. Ángel del Blanco1, Javier Torrente1, Pablo Moreno-Ger1, BaltasarFernández-Manjónl [5] titled as "Enhancing Adaptive Learning and Assessment in Virtual Learning Environments with Educational Games" The rising acceptance of Virtual Learning Environments (VLE) in the e-Learning field poses new challenges such as producing student-centered courses which can be automatically tailored to each student's needs. Morten Flate Paulsen[6] titled as "Online Education Systems: Discussion and Definition of Terms"This paper is written in order to establish a common framework of terms for the Web Education Systems Project (Web-edu), which is supported by the European Leonardo da vinci program.

2. MATERIALS AND METHODS

This study collected a data of 300 College students at different areas in and around Theni and Dindigul district and found that students having their own computers with internet access during their high school, and considering themselves to be intermediate with computers and this indicates the growth of e-learning .Therefore, the survey can claim to give a more comprehensive view about e-learning in the Theni District and its rate of development than has been available up to now. The data's are collected from students in though questionnaires and by personal interviews.

Description of Data Collection

Student's data are collected with different majors as Computer Science (CS), Biology (Bio), Management Studies (BBA). The attributes of collected data are,

1. Higher secondary Mark.
2. Gender.
3. Father Occupation
4. Mother Occupation
5. Rural & Urban Area
6. Medium
7. Practical Knowledge – Traditional & Virtual.
8. Learning Mark - Traditional & Virtual.
9. Assignment - Traditional & Virtual.
10. Seminar - Traditional & Virtual.

Table 1: Sample Dataset

S.NO	GEN	MED	AR	HS	MAJ	F-OCCU	M-OCC	Mark		Practical		Assignment		Seminar	
								T-M	V-M	T-P	V-P	T-A	V-A	T-S	V-S
1	F	E	U	1100	CS	WORKING	WORKING	45	50	M	VG	G	VG	M	VG
2	M	E	U	900	CS	WORKING	HOME MAKER	30	43	B	G	M	G	M	VG
3	F	T	R	800	CS	WORKING	WORKING	41	48	B	VG	M	G	B	VG
4	F	E	R	990	CS	WORKING	WORKING	45	50	M	VG	G	VG	M	VG
5	F	T	R	700	CS	WORKING	WORKING	30	45	M	VG	M	VG	M	VG

Classification and Clustering

Classification: Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifier, predict categorical class labels. Many Classification methods have been proposed by researchers in machine learning, pattern recognition and statistics. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large amounts of risk resident data. **Prediction:** Identification of one thing based purely on the description of another related thing. Not necessarily future events, just unknowns based on the relationship between thing that know and a thing need to predict.

Classification and Prediction methods are compared and evaluated according to following criteria:

- **Accuracy:** The Accuracy of a classifier refers to the ability of a given classifier to recently predict the class label of new or previously unseen data. Similarly the accuracy of a predictor refers to how well a predictor can guess the value of the predicted attribute for new or previously used data. Accuracy can be estimated one or more test sets that are independent of the training set.
- **Speed:** It refers to the computational costs involved in generating and using the given classifier or predictor.
- **Robustness:** The ability of the classifier or predictor to make correct predictions given noisy data or data with missing values.
- **Scalability:** Refers to the ability to construct the classifier or predictor efficiently given large amount of data.
- **Interpretability:** Refers to the level of understanding and the insight that is provided by the classifier or predictor.

Classification Using Decision Tree

Decision Tree is a Classification scheme which generates a tree and a set of rules, representing the model of different classes, from a given data set. DT is a flow chart like tree structure, where each internal node denotes a test on an attributes, each brands represents an outcome of the test leaf nodes represent the classes or class distributions the top most node in a tree is a root node. Easily derive the rules corresponding to the tree by traversing each leaf of the tree starting from the node. It may be noted that many different leaves of the tree may refer to the same class labels, but each leaf refers to a different rule. The major strength of the DT methods is the following:

- DT are able to generate understandable rules
- They are able to handle both numerical and categorical attributes

Classification trees are also called Decision Trees are especially attractive in a data mining environment for several reasons.

- First due to their intuitive representation, the resulting Classification model is easy to assimilate by humans.

- Second, decision trees do not require any parameter setting from the user and thus are especially suited for exploratory knowledge discovery.
- Third, DT can be constructed relatively fast and the accuracy of DT is comparable or superior to other classification models.

J48 Tree

J48 a can be called as optimized implementation of the C4.5 or improved version of the C4.5.

The limitations of C4.5 are discussed below.

(a) Empty branches**:** Constructing tree with meaningful value is one of the crucial steps for rule generation by C4.5 algorithm. In our experiment, we have found many nodes with zero values or close to zero values. These values neither contribute to generate rules nor help to construct any class for classification task. Rather it makes the tree bigger and more complex.

(b) Insignificant branches**:** Numbers of selected discrete attributes create equal number of potential branches to build a decision tree. But all of them are not significant for classification task. These insignificant branches not only reduce the usability of decision

(c) Over fitting**:** Over fitting happens when algorithm model picks up data with uncommon characteristics. This cause many fragmentations is the process distribution. Statistically insignificant nodes with very few samples are known as fragmentations. Generally C4.5 algorithm constructs trees and grows it branches 'just deep enough to perfectly classify the training examples'. This strategy performs well with noise free data. But most of the time this approach over fits the training examples with noisy data. Currently there are two approaches are widely using to bypass this over-fitting in decision tree learning.Those are:

- ✓ If tree grows very large, stop it before it reaches maximal point of perfect classification of the training data
- ✓ Allow the tree to over-fit the training data then post-prune tree.

AD Tree

An alternating decision tree (ADTree) is a machine learning method for classification. It generalizes decision trees and has connections to boosting. An alternating decision tree consists of decision nodes and prediction nodes. Decision nodes specify a predicate condition. Prediction nodes contain a single number . ADTree always has prediction nodes as both root and leaves. An instance is classified by an ADTree by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed.

The inputs to the alternating decision tree algorithm are

- A set of inputs $(x_1,y_1)\ldots(x_m,y_m)$ where $x_i$is a vector of attributes and y_i is either -1 or 1. Inputs are also called instances.
- A set of weights w_i corresponding to each instance.

The fundamental element of the ADTree algorithm is the rule. A single rule consists of a precondition, a condition, and two scores. A condition is a predicate of the form "attribute <comparison> value." A precondition is simply a logical conjunction of conditions. Evaluation of a rule involves a pair of nested if statements:

```
if (precondition)
    if(condition)
    return score_one
    else
    return score_two
    end if
else
return 0
```

```
endif
```

RandomTree

Random Tree is a supervised Classifier; it is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In standard tree each node is split using the best split among all variables. In a random forest, each node is split using the best among the subset of predicators randomly chosen at that node. Random trees have been introduced by Leo Breiman and Adele Cutler. The algorithm can deal with both classification and regression problems. A random tree is a collection (ensemble) of tree predictors that is called forest. The classification works as follows: the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". In case of a regression, the classifier response is the average of the responses over all the trees in the forest.

The algorithm for the Random tree classification algorithm is given below:

```
Start
        {
                RF= {Choose attributes subset of given dataset D}
                For each chosen variable
                {
                        If (RF.av == True) then take the relevant attributes
                        Else
                        Take the unrelevant attributes
}
                        For all RF until leaf node is reached.
                        End   }
```

3. RESULTS AND DISCUSSION

The paper discusses and analyses about the results that obtained from the classification algorithms such as AD Tree, J48 Tree and Random Tree through the,

- Attribute Distribution
- Tree view of the Classifier algorithms
- Stratified Cross Validation Analysis
- Comparisons

Attribute Distribution

The attribute of virtual and Traditional dataset contains subset of 100 data from those randomly 25 instances of data and 8 attributes are Medium, Hs-Mark, Major, V-Practical-Mark, V-Assignment, V-Seminar, and Gender.

Fig 1 Attribute distribution of Virtual Dataset

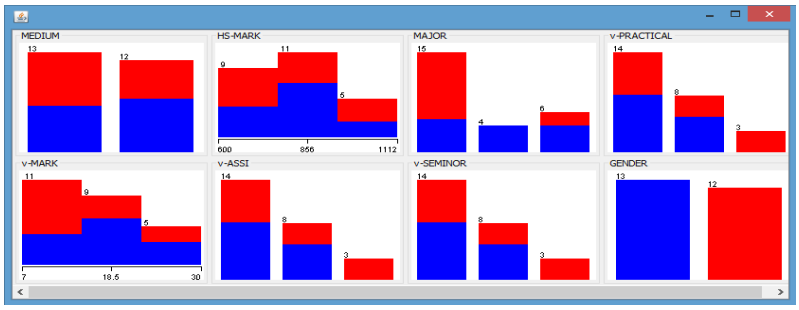
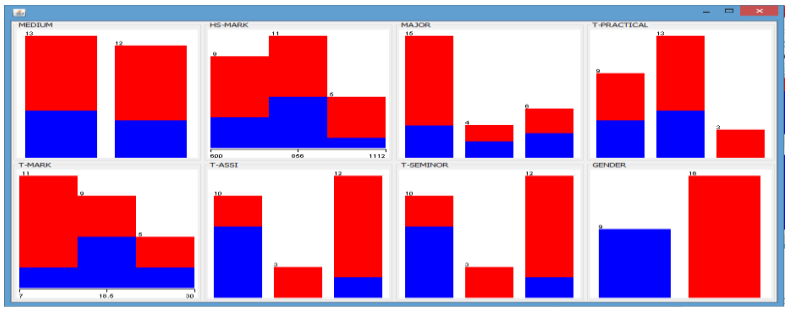


Fig 2 Attribute distribution of Traditional Dataset



Tree view and Accuracy of Classification Algorithm

The goal of the classification technique is to predict the target accurately for each case in the data. The important accuracy estimators are listed below:

- **TP Rate :** Rate of True Positives(instances correctly classified as a given class)
- **FP Rate :** Rate of False positives(instances falsely classified as a given class)
- **Precision :** Proportion of instances that are truly of a class divided by the total instances classified as that class
- **Recall :** Proportion of instances that are truly of a class divided by the actual total in that class(equivalent to TP rate)
- **F-Measure :** A combined measure for precision and recall calculated as **2*Precision*Recall/(Precision + Recall)**

The accuracy estimator of ADTree, J48 Tree and Random algorithm for the traditional dataset and virtual dataset are listed on the below tables.

Table 2 Classification accuracy of ADTree

Type of Data	# of instances	TP Rate	FP Rate	Precision	Recall	F-Measure	Class Label
Virtual Dataset	25	1	0	1	1	1	M
		1	0	1	1	1	F
Traditional Dataset	25	1	0	1	1	1	M
		1	0	1	1	1	F

Table 3 Classification accuracy of J48 Tree

Type of Data	# of instances	TP Rate	FP Rate	Precision	Recall	F-Measure	Class Label
Virtual Dataset	25	0.923	0.333	0.75	0.923	0.828	M
		0.667	0.077	0.889	0.667	0.762	F
Traditional Dataset	25	0.778	0.063	0.875	0.778	0.824	M
		0.938	0.222	0.882	0.938	0.909	F

Table 4 Classification accuracy of Random Tree

Type of Data	# of instances	TP Rate	FP Rate	Precision	Recall	F-Measure	Class Label
Virtual Dataset	25	1	0	1	1	1	M
		1	0	1	1	1	F
Traditional Dataset	25	1	0	1	1	1	M
		1	0	1	1	1	F

By comparing all the three algorithms with the two virtual and traditional datasets, it is observed that the both datasets are well classified in ADTree and Random Tree algorithms.

Stratified Cross- Validation Analysis

The percentage of correctly classified instances is often called "accuracy" or "simple accuracy". The error rates are used for numeric prediction rather than classification. The analysis involves lots of error estimators like Mean Absolute Error (MAE). Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) and Kappa Statics'.

Kappa is a chance-corrected measure of agreement between classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. The Kappa value greater than zeros shows that the classifiers doing better chance. To predict the error estimators like MAE,RMSE,RAE,RRSE and Kappa statistics, in the process of classification task , a simulation trails were conducted.

The error simulation trails conducted for virtual learning and traditional learning with AD Tree, J48 Tree and Random Tree algorithms are analysed in the table 5 and compared in the chart in fig 3.

Table 5 Error estimator of the simulation trials to VL

Algorithm	Instance	Kappa	MAE	RMSE	RAE(%)	RRSE(%)	Correctly classified	Incorrectly classified
ADTree	25	1	0.1971	0.237	39.4773%	47.4321%	25	0
J48		0.5955	0.2667	0.3651	53.4125%	73.0878%	20	5
RandomTree		1	0	0	0%	0%	25	0

Fig 3 Bar chart comparison between error estimators in VL

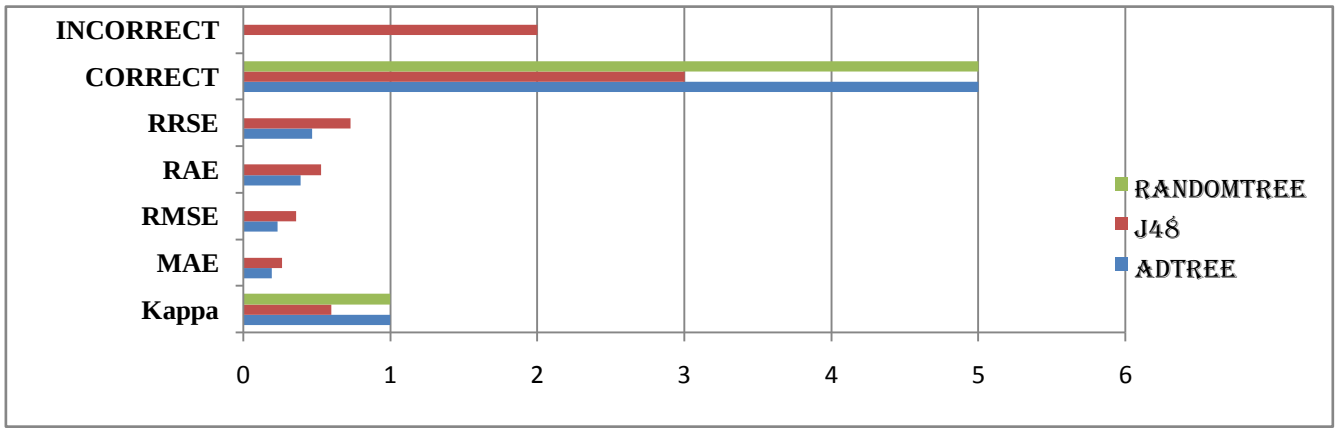
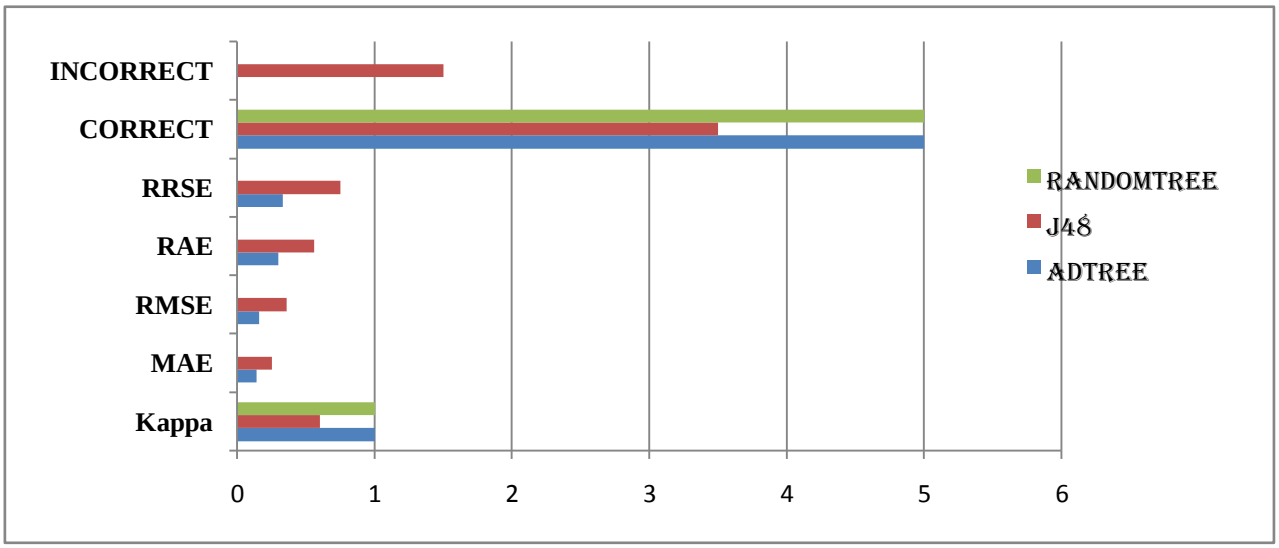


Table 6 Error estimator of the simulation trials to TL

Algorithm	Instance	Kappa	MAE	RMSE	RAE(%)	RRSE(%)	Correctly classified	Incorrectly classified
ADTree	25	1	0.1402	0.1606	30.2394%	33.4561%	25	0
J48		0.6528	0.2598	0.3604	56.0334%	75.0734%	21	4
RandomTree		1	0	0	0%	0%	25	0

Fig 4 Bar chart comparison between error estimators in TL



The tabulated values are plotted as a bar chart. From the table and the chart, it's observed that among all the datasets, the AD Tree and Random Tree have lower error estimators and also it has accurate number of correctly classified instances and getting lower error rate signals that the classifier used in this study works good. The kappa value is greater than zero, in ADTree and Random Tree and it classifies well.

4. Conclusion

The result of all the above algorithms can be analyzed through the two virtual and traditional learning datasets from that the ADTree and Random Tree shows better accuracy , inorder to find which learning is best we move on to manual comparison. Through all these test it's concluded the virtual learning is best for the upcoming students to understand their subjects clearly.

5.REFERENCES

[1] Mary C Dyson and Silvio BarretoCampello, "Evaluating Virtual Learning Environments: what are we measuring?", Electronic Journal of e-Learning, Volume 1 Issue 1 (2003) 11-20

[2] Trilok Chand Sharma and Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm",International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013

[3] Lukman, R., and KrajncM,"Exploring Non-traditional Learning Methods in Virtual and Real-world Environments".

[4] Peter W. Stonebraker and James E. Hazeltine, "Virtual learning effectiveness An examination of the process",Northeastern Illinois University, Chicago, Illinois, USA

[5] Angel del Blanco1, Javier Torrente1, Pablo Moreno-GerlandBaltasarFernández-Manjónl ,"Enhancing Adaptive Learning and Assessment in Virtual Learning Environments with Educational Games",UniversidadComplutense de Madrid. Facultad de Informática.C/ Profesor José GarcíaSantesmases s/n. 28040 Madrid, Spain.

[6] Morten Flate Paulsen titled as "Online Education Systems: Discussion and Definition of Terms"

[7] https://en.wikipedia.org/wiki/virtual_learning.