Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms

T. Jeevalatha
M. Phil Scholar, Department of Computer Science,
Dr. N.G.P Arts and Science College, Coimbatore,
Tamil Nadu, India.

N. Ananthi
Assistant Professor,
Department of Computer Technology,
Dr. N.G.P Arts and Science College, Coimbatore, TN, India

D. Saravana Kumar
Assistant Professor and Placement Officer,
Adithiya Institute of Technology, Coimbatore,
TN, India.

ABSTRACT

Data mining is a new approach for education. The main objectives of higher education institutions are to provide quality education to its students for their better placement opportunity. We could use Decision tree algorithms to predict student selection in placement. It helps us to identify the dropouts of the student who need special attention and allow the teacher to provide appropriate placement training. This paper describes how the different Decision tree algorithms used to predict student performance in placement. In the first step we have gathered the last two years passed out students details from placement cell in Dr.N.G.P Arts and Science College. In the second step preprocessing was done on those data and attributes were selected for prediction and in the third step Decision tree algorithms such as ID3, CHAID, and C4.5 were implemented by using Rapid Miner tool. Validation is checked for the three algorithms and accuracy is found for them. The best algorithm based on the collected placement data is ID3 with an accuracy of 95.33%.

General Terms

Data mining, classification, Decision tree algorithms.

Keywords

Data mining, Decision tree, CHAID, placement prediction, C4.5, ID3

1. INTRODUCTION

The campus placement of the students plays an important role in an educational institution. The companies identify the talented and qualified professionals before they completed their education. So the major success of institution is giving the placement chance to the students. The main motive of this paper is to classify the placement of candidates by using decision tree algorithms and rapid miner tool.

Data mining , database applications that look for patterns that are hidden (the gathering the information from already existing data stored in a database) in a bunch of data that can be used to determine future behaviors. The main motive of the data mining process is to take information from a data set and transform it into meaningful information for further analysis [1]. Data mining have a various types of techniques. The common data mining tasks are predictive model and descriptive model. A predictive model makes an excepted about data values using known result found from various data. Predictive modeling may be made based on the use of historical data. A descriptive model identifies relationships or patterns in data.

Classification a data mining technique which is most commonly used, it employs a set of data which are pre-classified. The classification process involves learning and classification. In learning by using classification algorithm, the training data were analyzed. In classification the accuracy of the classification rules are estimated by using test data [2].

Decision trees algorithms, tree-shaped structures that represent decision sets. They generate rules, which are used for the classification of data. Decision trees are the supportable technique for building understandable models. Once the tree is built, it is applied to each and every tuple in the database and result in a classification for those tuples [3].

Data mining in standard education is a recent area of research and this field of research is earning at higher rates because of its developments and progressive in educational institutes. Data Mining can be used in educational field to enhance the understanding of learning process to focus on finding, extracting and validating variables related to the student learning process. Mining in educational environment is called Educational Data Mining [2].

Knowing the factors for placement of student can help the teachers and administrators to take necessary actions so that the success percentage of placement can be improved. Predicting the placement of a student needs a lot of parameters that are to be considered. Prediction of models includes all personal, social, and psychological and other variables are required for the effective prediction of the placement of the student.

2. LITERATURE REVIEW

In this paper[4], they have used the Artificial Neural Network and Decision tree(C5) algorithms and they collected high school mark, higher secondary mark and category(arts or science) ,degree mark in last semester and student basic information (Name, roll no, sex, department).The collected data was processed in artificial neural network and decision tree to predict who are getting chance to be placed in the interview and they concluded that decision tree is the best algorithm and its accuracy is 95%. This work is helpful to improve the performance of weak students in early stage. In this paper [5], they collected student's previous and current academic data from engineering institutions .The collected data are 10th, 12th, B.Tech passing percentage and other information, these data were applied into decision tree. The decision tree generated the various rules and finally the output is predicted as whether the students are placed or not in the campus interviews.

Ajay Kumar Pal et al [6], applied the decision tree model to predict the student' performance in placement. They collected the data like MCA mark, lab work, , communication skill. The data were applied into different classification algorithms (Decision tree, neural network and Naïve Bayes). The algorithms were being preceded in data mining tool – WEKA. The result is used to identify which algorithm is best to predict student's performance in placement and the result is Naïve Bayes classifier and the accuracy is 86.15%.

In this paper [7], obtained the model to classify student placement performance. They collected the 325 sample data in various institutions in Mumbai. The data are SSLC, HSC, UG percentage and PG final semester result. They took 195 records for training the predicting student's performance in placement and remaining 130 records used to validate the model. The data are applied into decision tree algorithm and they have used dot net software for classifying the student's placement performance in colleges.

In this paper [8], they have used the decision tree learning algorithms such as ID3, ASSISTANT and C4.5. They collected enrolled students data from engineering institutes that have different information about their previous and current academic records like student roll no, name, date of birth, 10th, 12th, undergraduate passing percentage and applying these details in decision tree method for classifying student academic performance for placement department and they concluded that data mining learning techniques are helpful for placement prediction in colleges.

Hence the research goal of the study is to use data mining algorithms and various tools to predict the performance and placements of student based on previous academic record. In our report, we are giving the various existing Decision tree algorithms present for classification and which is the best for the prediction of placement selection in students data set.

3. DECISION TREE ALGORITHMS

3.1 ID3 (Iterative Dichotomiser)

It works on the principle of the Occam's razor and used to create the smallest possible decision tree. It takes all the attributes which are unused and promotes the calculation of entropies which are used to measure the informative of node. It also scans and chooses the attribute which has the entropy is less or when information gain is large [8].

3.2 C4.5 (Successor of ID3)

C4.5 is a popular algorithm which is used for the generation of a decision trees. It is an advanced level of the ID3 algorithm which is designed to overcome its limitations. The decision trees generated by this algorithm are used for prediction and it is a classifier of statistical type [9].

3.3 CART

CART is the acronym of Classification And Regression Tree. It is one of the well known methods of building [10]. CART builds a tree which is generally binary decision tree by breaking up the records at each node, in according to a functional work of every single attribute for determining a best split, it uses the gini index criteria.

3.4 CHAID

CHAID is the short version of CHi-squared Automatic Interaction Detector. When computing classification trees it is used for performing multi-level splits. CHAID is usually used for prediction and also classification, and even for detection of interaction between different variables [11]. The main advantage of CHAID over the others is CHAID is non-parametric. CHAID tries to stop growing the tree before the occurring of over fitting.

4. TECHNOLOGY USED

Rapid Miner is an open source data mining tools that provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. Rapid Miner provides 99% of an advanced analytical solution through template-based frameworks that speed delivery and reduce errors by nearly eliminating the need to write code [12].

5. ARCHITECTURE OF WORK

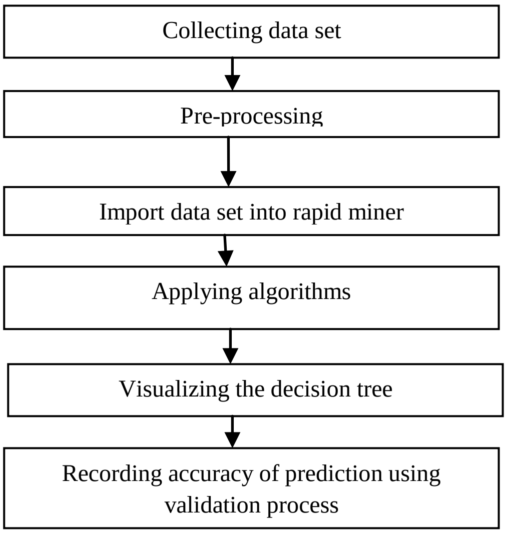

Collecting data set
Pre-processing
Import data set into rapid miner
Applying algorithms
Visualizing the decision tree
Recording accuracy of prediction using validation process

Figure 1: Architecture of work

6. DATA COLEECTION

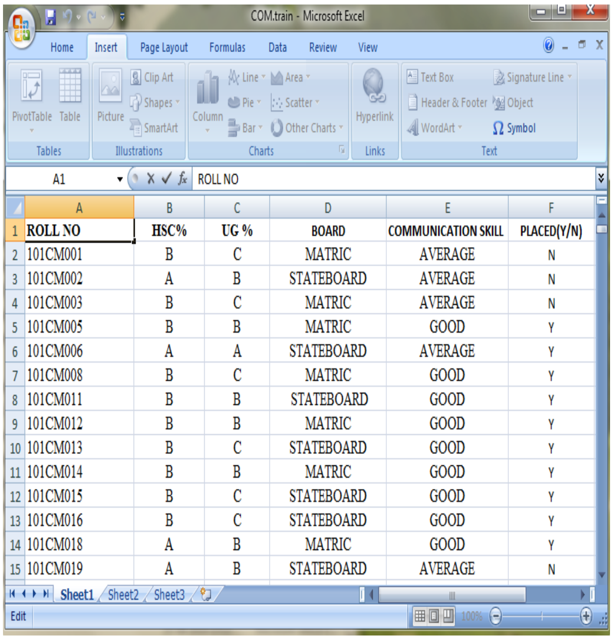
Table 1: The definition of attributes and values

Attribute	Definition	Values
Roll number	Identity of a student	ROLL NO
HSC	Higher secondary board marks	A>80 B>60 C<60
UG	Under graduate marks	A>80 B>60 C<60
Board	Medium of education in HSC	MATRIC/STATE BOARD
Communication	Way of conveying	GOOD/AVERAGE
Placed(y/n)	Details about selection in campus interview	Y/N

7. DATA PREPROCESSING

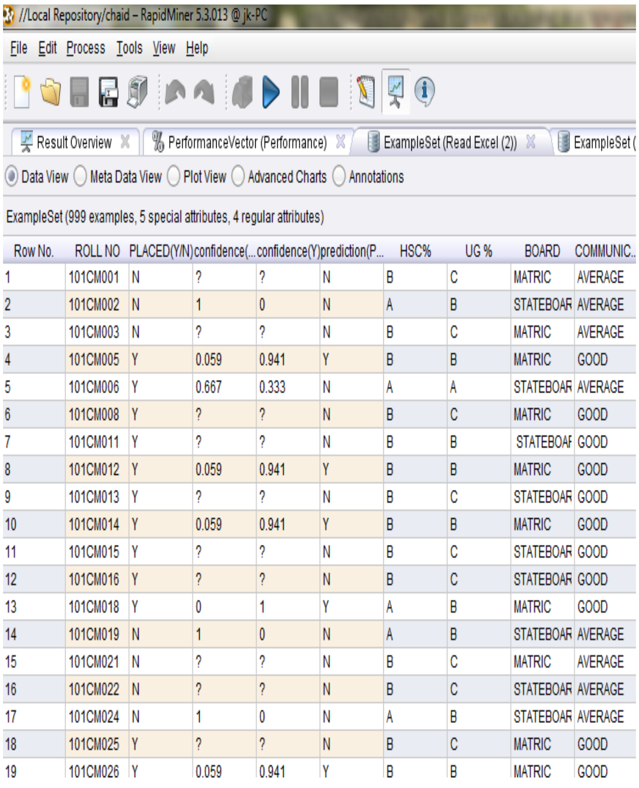
Raw data is a quality less and inconvenience data for processing. This poor quality of raw data affects the data mining efficiency. In order to improve the quality of the data and, also the mining results pre-processing of raw data is carried out. Data preparation and filtering steps takes large amount of time [11].

In our case, we have collected 1342 students details as a dataset from placement cell in Dr.N.G.P arts and Science College and the missing values were collected from the respective departments and tabulated. Here, irrelevant attributes such as students residential address, name, etc had been removed. Since the algorithms such as ID3, CHAID and C4.5 supports only nominal values, the values of the "HSC" and "UG" attribute of the students were converted as "A" if the percentage of marks scored by the student is 80 and above, "B" if the percentage was less than 60 and greater than 80, "C" if the percentage was less than 60. And also their board of education is either matriculation or state board.



	A	B	C	D	E	F
1	ROLL NO	HSC%	UG %	BOARD	COMMUNICATION SKILL	PLACED(Y/N)
2	101CM001	B	C	MATRIC	AVERAGE	N
3	101CM002	A	B	STATEBOARD	AVERAGE	N
4	101CM003	B	C	MATRIC	AVERAGE	N
5	101CM005	B	B	MATRIC	GOOD	Y
6	101CM006	A	A	STATEBOARD	AVERAGE	Y
7	101CM008	B	C	MATRIC	GOOD	Y
8	101CM011	B	B	STATEBOARD	GOOD	Y
9	101CM012	B	B	MATRIC	GOOD	Y
10	101CM013	B	C	STATEBOARD	GOOD	Y
11	101CM014	B	B	MATRIC	GOOD	Y
12	101CM015	B	C	STATEBOARD	GOOD	Y
13	101CM016	B	C	STATEBOARD	GOOD	Y
14	101CM018	A	B	MATRIC	GOOD	Y
15	101CM019	A	B	STATEBOARD	AVERAGE	N

Figure 2: Example Dataset in Excel



ExampleSet (999 examples, 5 special attributes, 4 regular attributes)

Row No.	ROLL NO	PLACED(Y/N)	confidence(...	confidence(Y)	prediction(P...	HSC%	UG %	BOARD	COMMUNIC...
1	101CM001	N	?	?	N	B	C	MATRIC	AVERAGE
2	101CM002	N	1	0	N	A	B	STATEBOAR	AVERAGE
3	101CM003	N	?	?	N	B	C	MATRIC	AVERAGE
4	101CM005	Y	0.059	0.941	Y	B	B	MATRIC	GOOD
5	101CM006	Y	0.667	0.333	N	A	A	STATEBOAR	AVERAGE
6	101CM008	Y	?	?	N	B	C	MATRIC	GOOD
7	101CM011	Y	?	?	N	B	B	STATEBOAF	GOOD
8	101CM012	Y	0.059	0.941	Y	B	B	MATRIC	GOOD
9	101CM013	Y	?	?	N	B	C	STATEBOAR	GOOD
10	101CM014	Y	0.059	0.941	Y	B	B	MATRIC	GOOD
11	101CM015	Y	?	?	N	B	C	STATEBOAR	GOOD
12	101CM016	Y	?	?	N	B	C	STATEBOAR	GOOD
13	101CM018	Y	0	1	Y	A	B	MATRIC	GOOD
14	101CM019	N	1	0	N	A	B	STATEBOAR	AVERAGE
15	101CM021	N	?	?	N	B	C	MATRIC	AVERAGE
16	101CM022	N	?	?	N	B	C	STATEBOAR	AVERAGE
17	101CM024	N	1	0	N	A	B	STATEBOAR	AVERAGE
18	101CM025	Y	?	?	N	B	C	MATRIC	GOOD
19	101CM026	Y	0.059	0.941	Y	B	B	MATRIC	GOOD

Figure 3: Output of the training dataset in CHAID

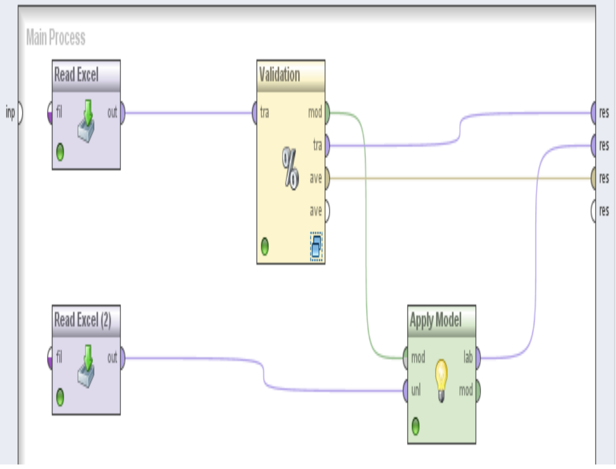
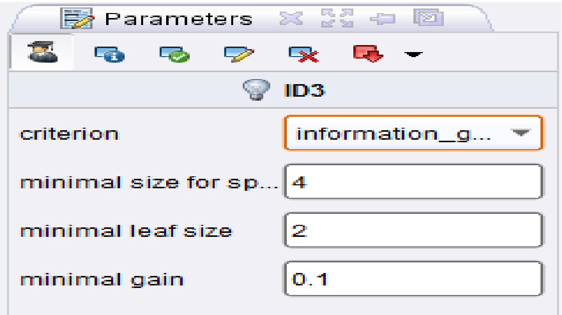


Figure 4: Design of the validation process

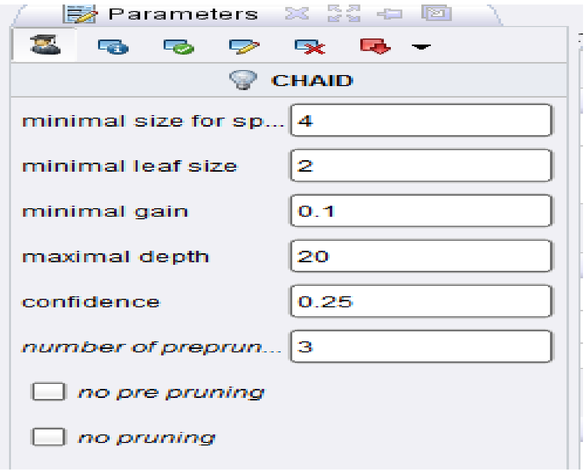
7.1 Working with Algorithms

The operator cannot handle the numerical attributes. We can import the excel dataset(Figure 2) by using *Read Excel* operator and change the attribute type such as Roll number as ID and the Placed(y/n) as class label. Connect the *out* port to the *tra* port in *ID3 algorithm* operator and change the criterion as information gain (Figure 5)and connect the *mod* with *res* port. After that click the run button to see the tree structure for ID3 algorithm(Figure 7) and output for the training dataset(Figure 3). For validation, in the design process the *validation* operator should be added for measuring the accuracy, precision and recall values(Figure 4) and click on the run button to see the confusion matrix of the algorithm. It was similar for C4.5 and CHAID algorithms(Figure 6).



criterion	information_g...
minimal size for sp...	4
minimal leaf size	2
minimal gain	0.1

Figure 5: Parameter for ID3 algorithm



minimal size for sp...	4
minimal leaf size	2
minimal gain	0.1
maximal depth	20
confidence	0.25
number of preprun...	3

□ *no pre pruning*

□ *no pruning*

Figure 6: Parameter for CHAID and C4.5 algorithm

8. EXPERIMENTAL RESULTS

Graph view of the decision tree

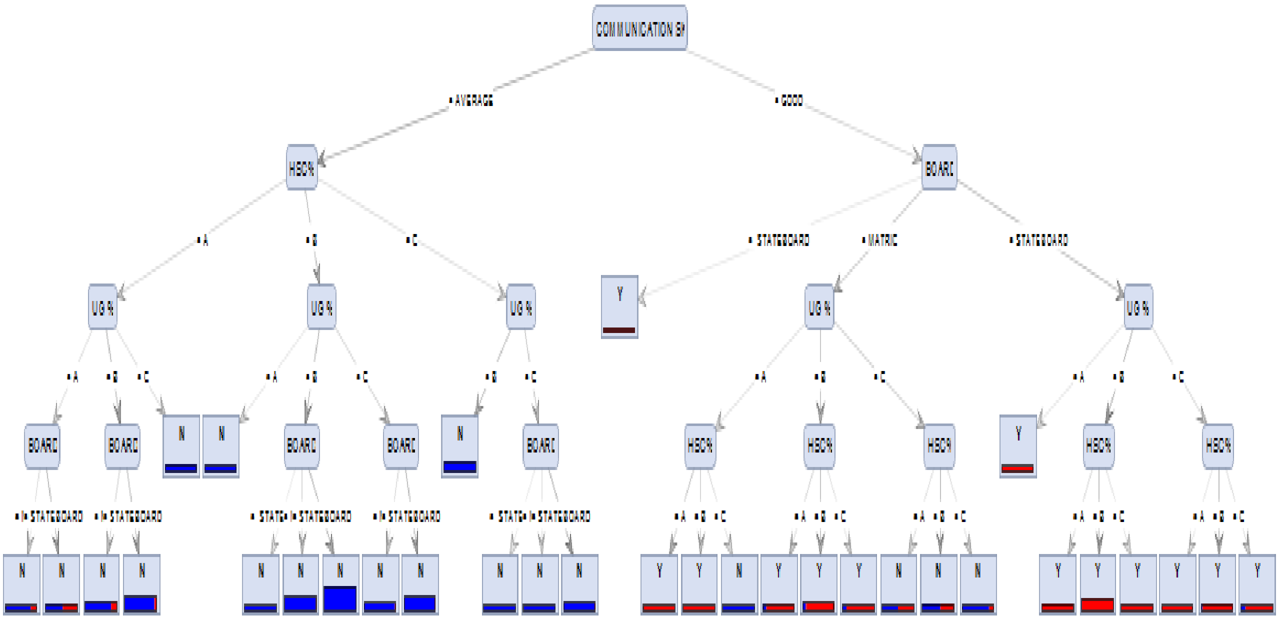
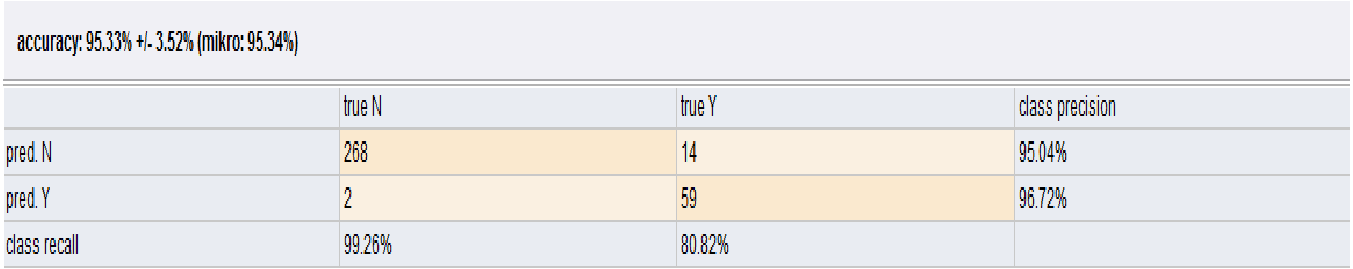


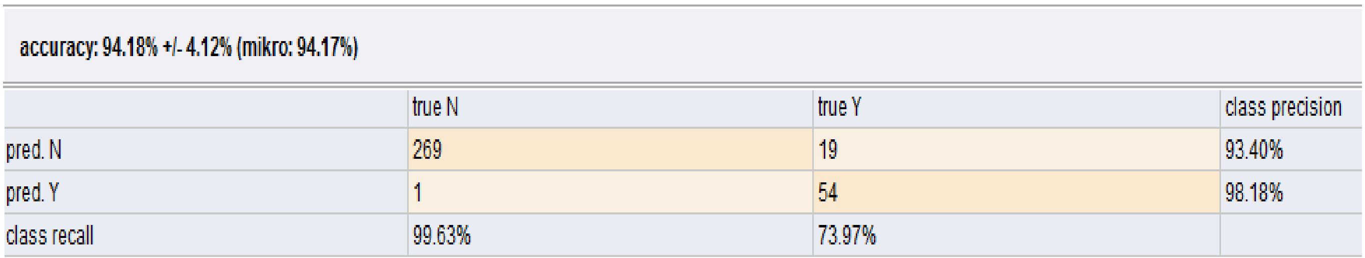
Figure 7: Tree structure for Decision Tree algorithm



accuracy: 95.33% +/- 3.52% (mikro: 95.34%)

	true N	true Y	class precision
pred. N	268	14	95.04%
pred. Y	2	59	96.72%
class recall	99.26%	80.82%	

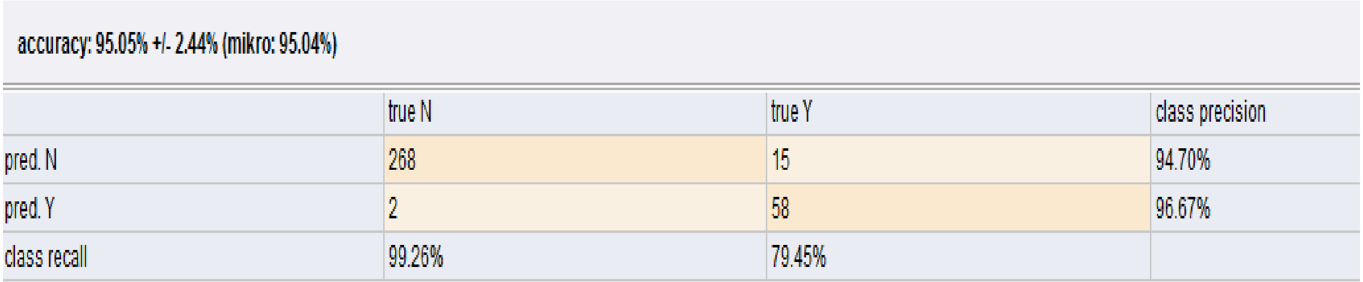
Figure 8: Confusion matrix for ID3 Algorithm



accuracy: 94.18% +/- 4.12% (mikro: 94.17%)

	true N	true Y	class precision
pred. N	269	19	93.40%
pred. Y	1	54	98.18%
class recall	99.63%	73.97%	

Figure 9: Confusion matrix for CHAID Algorithm



accuracy: 95.05% +/- 2.44% (mikro: 95.04%)

	true N	true Y	class precision
pred. N	268	15	94.70%
pred. Y	2	58	96.67%
class recall	99.26%	79.45%	

Figure 10: Confusion matrix for C4.5 Algorithm

Precision and recall:
Precision: Percentage of selected items that is correct

Recall: Percentage of correct items that are selected

$$\text{Precision} = \frac{\text{true positive}}{\text{True positive + false negative}}$$

$$\text{Recall} = \frac{\text{true positive}}{\text{True positive + false negative}}$$

$$\text{Accuracy} = \frac{\text{true positive + true negative}}{\text{True positive + false positive + false negative + true negative}}$$

Table 2: Accuracy of the algorithms

Algorithm	**Total number of Students**	**Students whose result was wrongly predicted**	**Accuracy**
ID3	1342	68	95.33
C4.5	1342	81	95.05
CHAID	1342	90	94.18

9. CONCLUSION

The C4.5, ID3 and CHAID data mining techniques were implemented on students data. From the results obtained above it is proved that Decision tree algorithms are applied on students data for analyzing the students placement selection. The efficiency of different decision tree algorithms was analyzed based on the accuracy (Figure 8, Figure 9, Figure 10). From the results it is found that ID3 algorithm is appropriate for predicting student placement. ID3 gives 95.33% prediction which is higher than C4.5 and CHAID algorithm (Table 2).

10. SCOPE FOR FUTURE WORK

The present research will have been planned to extend by using neural network algorithms, genetic algorithms and support vector machines and to add aptitude skill and technical skill as attributes in the dataset. And also we will collect dataset of final year students till four semesters and applying various classification algorithms to predict the placement of students before final semester.

11. ACKNOWLEDGEMENT

We express our sincere gratitude to the staffs and the management of Dr.N.G.P Arts and Science College for their guidance and support.

12. REFERENCES

[1] Romero, C., Ventura, S. and Garcia, E., "Data mining in Course management systems: Moodle case study and Tutorial". Computers & Education, Vol. 51, No. 1.pp.368- 384. 2008.

[2] Machado, L. and Becker, K. "Distance Education: A Web Usage Mining Case Study for the Evaluation of Learning Sites". Third IEEE International Conference on Advance Learning Technologies (ICALT'03), 2003.

[3] Moscow, J and Beck, J., "Some useful tactics to modify, Map and mine data from intelligent tutors". Natural Language Engineering 12(2), 195- 208. 2006.

[4] Bahasen, "Predicting and analyzing secondary education placement: A data mining approach", International journal of Expert system with applications, 2012, vol: 3 issue: 10, pgno: 9468-9476.

[5] Samrat Singh, Dr. Vikesh Kumar" Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education", International Journal of Computer Science and Network (IJCSN), Volume 1, Issue 4, August 2012.

[6] Ajay Kumar Pal, Saurabh Pal "Classification Model of Prediction for Placement of Students", I.J.Modern Education and Computer Science, 2013, 11, 49-56Published Online November 2013 in MECS.

[7] NeelamNaik, SeemaPurohit"Prediction of Final Result and Placement of Students using Classification Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 56– No.12, October 2012

[8] Samrat Singh, Dr. Vikesh Kumar "Performance Analysis of Engineering Students for Recruitment Using classification Data Mining Techniques" , | IJCSET |February 2013| Vol 3, Issue 2, 31-37.

[9] DS Kumar, N.Ananthi, M.Devi "An Approach to Automation Selection of Decision Tree based on Training Data Set", International Journal of Computer Applications(0975-8887), Volume 64-No.21, February 2013.

[10] Jaiwei Han, MichelinneKanber "Data mining concepts and techniques"

[11] V.Ramesh, P.Parkavi, P.Yasodha" Performance Analysis of Data Mining Techniques for Placement Chance Prediction" International Journal of Scientific & Engineering Research Volume 2, Issue 8, August-2011 1 ISSN 2229

[12] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao "PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol3, No.5, September 2013.