



Comparative Analysis of Classification Techniques for Predicting Computer Engineering Students' Academic Performance

Williamjeet Singh

Department of computer engineering
Punjabi University Patiala, India
Williamjeet@gmail.com

Prabhjot Kaur

Department of computer engineering
Punjabi University Patiala, India
Dhaliwaljot6@gmail.com

Abstract—The quality of education system affects its every country's growth. The top level of quality in the education system is achieved by extracting useful information for predictions regarding students' success rate and factors affecting student performance. This useful information is masked in the educational data set and is discovered by using data mining techniques. An early prediction of student performance helps authorities to provide extra coaching and counseling to increase the success rate. In this paper different classification techniques have been used to construct student SGPA prediction model based on student's social conditions and previous academic performance. Two algorithms REP Tree and J48 have been exercised on the 236 records of computer engineering students of Punjabi University to predict the third-semester performance of the students. J48 gives more accurate results than REP Tree for student performance prediction. The overall accuracy of J48 is 67.37% and for REP Tree is 56.78%.

Keywords— data mining; educational data; prediction; classification.

I. INTRODUCTION

Education provided at school level just increases country's literacy rate but the education provided at engineering institutes has the direct impact the economy and development of a nation.

A large number of engineering institutes have been set up across the India. However, the quality of education provided by engineering institutes depends upon the success rate of students. Prediction of student academic performance helps in identifying weak student. Thus, the management of engineering institute takes essential steps to improve weak student's performance.

Clustering, classification and association rule mining are data mining techniques which are used to extract knowledge from educational data set. This paper explores the affect of social parameters on student academic performance. Social parameters like early life, father's education, mother's education and present living scenario are chosen from the previous research done in the area of EDM.

The objective of data mining is to define the kind of knowledge. There are two categories of data mining tasks descriptive and predictive. Association rule mining and clustering are the descriptive data mining techniques used to extract hidden patterns from large data sets. Classification is predictive data mining technique used for prediction of a class of new data set.

There are mainly two types of classification- Black Box and White Box classification. White Box classification algorithm constructs models which provide the output in the form of IF-THEN rules, which are easy to interpret. White Box classification is used directly for intelligent decision making. The Black Box classification is more accurate but difficult to interpret.

The objective of this paper is to predict SGPA of B. Tech third-semester computer engineering students. The reason to consider the third-semester for SGPA prediction is the observation that some students drop out after the first year, some students change their stream and students start learning all the computer related subjects in third semester. Decision tree algorithm: J48 and REP Tree have been used to construct the model and finding an impact of social parameters on student academic performance.

This paper organized in this manner- Section II presents previous work done in EDM. Section III provides experimental settings followed by results in Section IV and conclusions are discussed in Section V.

II. LITERATURE SURVEY

H. Guruler et al.[32] developed a system MUSKUP to analyze performance of new student. Classification technique is applied on the student data to find demographic data which affects student GPA most. Their results showed the income levels of the students' family and the types of registration to the university were associated with student success.

Kishore, Venkatramaphanikumar and Alekhya[13] considered 9 attributes to the predict performance of computer science engineering students at Vignan University. Data set of 60 students is used as training data and testing can be done on data set of 134 students. MLP has achieved highest accuracy among J48, Naïve Bayes, CART, RBF.

J. Gamulin et al.[33] compared the accuracy of classification techniques on 2-class model and 3-class model. They collected data set of 302 students enrolled in physics course at University of Zagreb School of Medicine using Moodle system and Pitalica tool. The results of their study showed that 2 classes model has higher accuracy than 3 classes model.

V. Ramesh et al.[31] analyzed the performance of data mining techniques for placement chance prediction. WEKA tool was used by them for implementation. They used 5 algorithms with accuracy given as following: NaïveBayes

Simple(83.193%), Multilayer Perception (87.395%), SMO (84.0336%), J48(84.8739%), REP Tree (84.8739%).

Mishra, Kumar and Gupta [7] attempted a study to compare performance of J48 and Random Tree algorithms to predict third-semester MCA students' performance. They applied decision tree algorithms on data set of 250 students with 25 attributes using WEKA. They found Random tree(94.418%) to be more accurate than J48(88.372%) algorithm for student performance prediction.

K. Bunker et al. [6] applied classification algorithm (CART, C4.5 and ID3) to predict performance of B.A. first year students at Vikram University, Ujjain. They build an interface that provides the use of generated rules to predict the final grades of students in a course under study.

S. Fong et al. [8] proposed a hybrid model of Neural Network and Decision Tree algorithm implemented using WEKA that predicts the university to which a student may get admission based on his academic merits, background and the university admission criteria. The data set of 2400 secondary school students in Macau was collected for this study.

Pradeep, Das and Kizhekkattam [9] analyzed the factors affecting students' performance. Data mining techniques were applied to data set of 670 students with 57 attributes using WEKA. Prims algorithm has the highest accuracy in this study with both 57 attributes and 12 best attributes among JRip, OneR, ADTree, J48 and Simplecart.

S.Taruna and M.Pandey[10] compared the five classification algorithms- Bayesian Network, K-Nearest Neighbor, Naïve Bayes Tree, Decision Tree and Naïve Bayes for prediction of engineering students grades. They classify student marks in four classes A, B, C and F. Bootstrap method available in WEKA was used to improve the accuracy of each classifier. IBK, Bayes Net and Decision Tree gave excellent results but the results given by Naïve Bayes and Naïve Bayes Tree were not satisfactory.

Kumar and Vijayalakshmi[19] attempted a comparative study of Decision Tree and OneR algorithms for student academic records evaluation in higher education. According to their results, OneR algorithm is more accurate than Decision Tree algorithm.

Z. Zakaria et al.[11] performed a research to find the link between the previous academic performance and teaching aid environment, gender, teaching methodology, lecture involvement and students' attitude of electrical engineering students in University Teknologi MARA, Malaysia. They collected the data of 90 newly enrolled students via questionnaires. Their result shows that male students get better grades than female students and environment factors influence student academic performance.

III. EXPERIMENTAL SETTINGS

The objective of the methodology used is to construct a classification model to classify a student's third-semester SGPA as POOR (<4.0), BAVG (4.0 to 6.9), AVG (7.0 to 7.9), GOOD (8.0 to 8.9), EXCLT (>=9.0). The methodology starts with data collection followed by preprocessing. Methodology ends with modeling and classification.

A. Data Collection

We have collected data from students pursuing B.Tech Computer Engineering from Department of Computer Engineering, Punjabi University, Patiala. The data collection was done via a structured questionnaire. A sample of 260 students having 17 attributes was collected, which includes social parameters and previous academic performance shown in TABLE I.

B. Data Preprocessing

The data collected from students was saved as comma separated values file. The cleaning process applied on data set includes eliminating missing values, identifying outliers, removing duplicates and correcting inconsistent values. Data sources from total 260 instances in raw data ended up in 236 instances after data cleaning process.

C. Modeling

The WEKA (Waikato Environment for Knowledge Analysis) tool is used for implementing classification algorithms. WEKA is an open source tool coded in JAVA. WEKA supports data mining techniques and machine learning algorithms. Algorithms available in WEKA can be directly applied to the data set.

D. Decision Tree

Classification of instances using Decision Tree based algorithms is done by arranging them from top node to the bottom node of the tree. Each node in the tree represents feature of the instance and branches declining from that node represents the possible values for that feature. REP Tree generates a tree based on reduced variance or information gain and pruned that tree by using reduced error pruning. J48 is the JAVA code of the C4.5 decision tree algorithm. Reduced error pruning is used in J48 algorithm for pruned tree. REP Tree and J48 generate both pruned and unpruned trees. Cross-validation method is used for testing of data set as it is suitable for smaller data set and provides best error estimation.

IV. RESULTS AND DISCUSSION

REP Tree and J48 were implemented on the data set using 10 fold cross validation. REP Tree summary is listed in Fig. 1 and rules obtained are in TABLE II. While the J48 algorithm summary is listed in Fig. 2 and the rules obtained are listed in TABLE III. The Performance of both the algorithms is evaluated based on precision, true positive (TP) rate and recall. True positive is defined as the number of positive values predictions which are actually positive. Recall is the number of actual positive values that are predicted positive. Precision is a number of positive values predicted that are actually positive. High recall means algorithm returns most of the relevant results while high precision indicates that results returned by the algorithm are more relevant than irrelevant. Comparison of REP Tree and J48 algorithms is shown in TABLE IV.

The conclusions made from the rules extracted from the REP Tree and J48 are

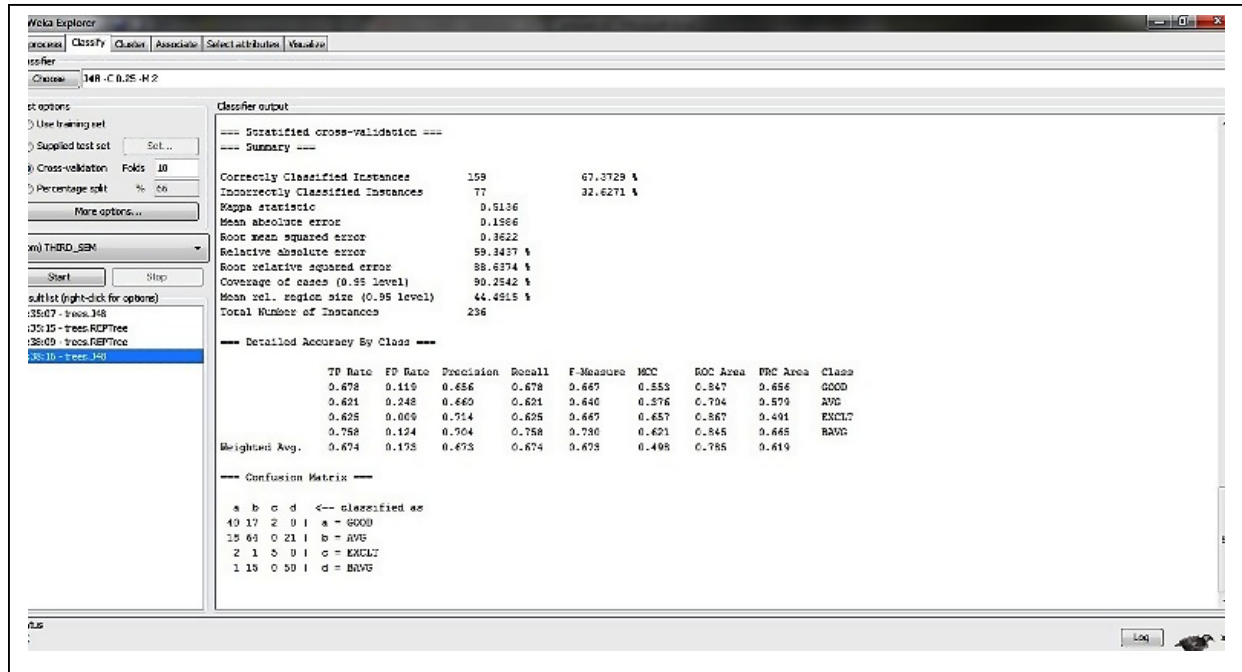


Fig. 3. J48 algorithm summary

TABLE III. RULES EXTRACTED FROM J48 ALGORITHM

1. If(SECOND_SEM=AVG)and (TOTAL10=GOOD): AVG
2. If(SECOND_SEM=AVG)and (TOTAL10=AVG): AVG
3. If(SECOND_SEM=AVG)and (TOTAL10=EXCLT)and (ELA=U): GOOD
4. If(SECOND_SEM=AVG)and (TOTAL10=EXCLT)and (ELA=R)and (ENGLISH10=EXCLT): AVG
5. If(SECOND_SEM=AVG)and (TOTAL10=EXCLT)and (ELA=R)and (ENGLISH10=GOOD): GOOD
6. If(SECOND_SEM=BAVG)and (TOTAL 10=EXCL): AVG
7. If(SECOND_SEM=BAVG)and (TOTAL 10=BAVG): BAVG
8. If(SECOND_SEM=BAVG)and (TOTAL 10=AVG)and (ME=P): AVG
9. If(SECOND_SEM=BAVG)and (TOTAL 10=AVG)and (ME=G)and (PHYSICS12=AVG): BAVG
10. If(SECOND_SEM=BAVG)and (TOTAL 10=AVG)and (ME=G)and (PHYSICS12=BAVG)and (ENGLISH12=GOOD): AVG
11. If(SECOND_SEM=BAVG)and (TOTAL 10=AVG)and (ME=G)and (PHYSICS12=BAVG)and (ENGLISH12=AVG): BAVG
12. If(SECOND_SEM=BAVG)and (TOTAL 10=AVG)and (ME=G)and (PHYSICS12=BAVG)and (ENGLISH12=BAVG): AVG
13. If(SECOND_SEM=BAVG)and (TOTAL 10=AVG)and (ME=N): BAVG
14. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST SEM=GOOD): AVG

15. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=AVG)and (TOTAL12=GOOD): BAVG
16. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=AVG)and (TOTAL12=AVG): AVG
17. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=AVG)and (TOTAL12=BAVG): AVG
18. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=BAVG)and (PHYSICS12=BAVG): BAVG
19. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=BAVG)and (PHYSICS12=GOOD): BAVG
20. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=BAVG)and (PHYSICS12=EXCLT)and (ENGLISH10=GOOD): BAVG
21. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=BAVG)and (PHYSICS12=EXCLT)and (ENGLISH10=AVG): AVG
22. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=BAVG)and (PHYSICS12=AVG)and (ME=P): AVG
23. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=BAVG)and (PHYSICS12=AVG)and (ME=N): AVG
24. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=BAVG)and (PHYSICS12=AVG)and (ME=G)and (SCIENCE10=EXCLT): AVG
25. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=BAVG)and (PHYSICS12=AVG)and (ME=G)and (SCIENCE10=GOOD): BAVG
26. If(SECOND_SEM=BAVG)and (TOTAL 10=GOOD)and (FIRST_SEM=BAVG)and (PHYSICS12=AVG)and (ME=G)and (SCIENCE10=AVG): BAVG
27. If(SECOND_SEM=EXCLT): GOOD
28. If(SECOND_SEM=GOOD)and (TOTAL10=AVG): AVG
29. If(SECOND_SEM=GOOD)and (TOTAL 10=GOOD)and (FIRST_SEM=AVG): AVG
30. If(SECOND_SEM=GOOD)and (TOTAL 10=GOOD)and (FIRST_SEM=GOOD): GOOD
31. If(SECOND_SEM=GOOD)and (TOTAL 10=EXCLT)and (FIRST_SEM=AVG): GOOD
32. If(SECOND_SEM=GOOD)and (TOTAL 10=EXCLT)and (FIRST_SEM=GOOD)and (ENGLISH10=EXCLT): EXCLT
33. If(SECOND_SEM=GOOD)and (TOTAL 10=EXCLT)and (FIRST_SEM=GOOD)and (ENGLISH10=GOOD): GOOD
34. If(SECOND_SEM=GOOD)and (TOTAL 10=EXCLT)and (FIRST_SEM=EXCLT): EXCLT

TABLE IV. PERFORMANCE COMPARISON OF REP TREE AND J48

	REP Tree			J48		
	TP Rate	Recall	Precision	TP Rate	Recall	Precision
GOOD	0.695	0.695	0.569	0.678	0.678	0.656
AVG	0.388	0.388	0.571	0.621	0.621	0.660
EXCL	0.000	0.000	0.000	0.625	0.625	0.714
BAVG	0.803	0.803	0.576	0.758	0.758	0.704
Weighted Average	0.568	0.568	0.673	0.674	0.674	0.674
Correctly Classified Instances	56.7797%			67.3729%		
Incorrectly Classified Instances	43.2203%			32.6271%		

The J48 (67.3729%) algorithm implementation attained higher accuracy than REP Tree(56.7797%) algorithm. The value of True Positive Rate, Recall and Precision measures of J48 algorithm are greater than REP Tree

V. CONCLUSION

Academic success of engineering students has become a major issue for the authorities. This study concentrates on identifying the attributes that affects student third-semester performance. This paper presents potential use of EDM using J48 and REP Tree algorithms to discover relationship between social parameters and student performance, and

predicting students' performances in third-semester. Analysis revealed that father's education and mother's education has affect on student performance and second-semester performance plays important role for third-semester performance. The results revealed that an early prediction of week students in academics helps the authorities to take necessary decisions for improving students' performance. J48 gave higher accuracy than REP Tree algorithm for student academic performance prediction. The future research will include prediction of B.Tech students' performance in all eight semesters and development of decision support system that helps authorities in identifying week students.

Education Placement-Test Scores: A Data Mining Approach," *Expert system with application*, vol. 39, no. 10, 2012.(refrences)

REFERENCES

- [1] Y.H.Yahaya,N.Wahab,M.R.M.Isa,N.F.Awang,H.Y.Seong M.Wook, "Predictechnique NDUM Student's Acadmic Performance Using Data Mining T," in *Second International Conference on Computer and Electrical Engineering*, 2009, pp. 357-361.
- [2] B. Sen,E. Ucar and D. Delen, "Predicting and analyzing Secondary
- [3] B.M. Bidgoli, D.Koshy, G.Kortemeyer,W.F.Punch, "predicting student performance: An applicant of data mining methods with an educational web based system," in *33rd ASEE/IEEE frontiers in Education Conference*, 2004.
- [4] M. Ramaswami and R. Bhaskaran, "A CHAID based performance prediction model in educational data mining," *international journal of computer sciences*, vol. 7, no. 1, 2010.

CONFERENCE PAPER

International Conference on

Recent Trends in Computer Science & Information Technology (RTCSIT-2016)

21st August 2016

Guru Nanak College Budhlada, Punjab India

- [5] T. Nghe, J. Paul, Aneek and Peter Heddawy, "A Comparative analysis of techniques for predicting academic performance," in *37th ASEE/IEEE frontiers in education conference*, 2007.
- [6] K. Bunker, R. Bunker, U. K. Singh, B. Pandya, "Data Mining: Prediction for Performance Improvement of graduate students using classification," in *9th international conference on wireless and optical communications network*, 2012, pp. 1-5.
- [7] T. Mishra, D. Kumar and S. Gupta, "Mining Students' data for performance prediction," in *4th International conference on advanced computing & communication technologies*, 2014, pp. 255-262.
- [8] S. Fong, Y.W.Si,R.P. Biuk-Aghai, "Applying a Hybrid Model of Neural Network and Decision Tree Classifier for Predicting University Admission," in *7th International conference on Information,communications and signal processing*, 2009, pp. 1-5.
- [9] A. Pradeep, S. Das, J.J. Kizhekkethottam, "Students Dropout Factor Prediction Using EDM techniques," in *International Conference on Soft-Computing and Network Security*, 2015, pp. 1-7.
- [10] S. Taruna and M. Pandey, "An Empirical Analysis of Classification Techniques for Predicting Academic Performance," in *IEEE international advanced computing conference*, 2014, pp. 523-528.
- [11] R.A.Kassim,A.Mohamad and N.Buniamin Z.Zakaria, "The Impact of Environment on Engineering Students' Academic Performance: A Pilot Study," in *3rd International Congress on Engineering Education*, 2011.
- [12] S.Huang and N.Fang, "Work in Progress-Prediction of Students' Academic Performance in an Introductory Engineering Course," in *41st ASEE/IEEE Frontiers in Education Conference*, Rapid City, 2011. (references)
- [13] V.S and S.Alekhyia K.V.K.Kishore, "Prediction of Student Academic Progression: A Case Study on Vignan University," in *International Conference on Computer Communication and Informatics*, Coimbatore,India, 2014. (references)
- [14] P. Pumpuang,A.Srivihok,P.Praneetpolgrang, "Comparisons of Classifier Algorithms: Bayesian Network,C4.5,Decision Forest and NBTree for Course Registration Planning Model of Undergraduate Students," in *IEEE International Conference on Systems,Man and Cybernetics*, Thailand, 2008. (references)
- [15] G. Kesavaraj andS. Sukumaran , "A study on classification techniques in data mining," in *Fourth international conference on computing,communications and networking technologies*, Tiruchengode, 2013, pp. 1-7. (references)
- [16] C. Romero and S. Ventura , "Data mining in education," vol. 2, no. 3, 2013.
- [17] C. Romero and S. Ventura , "Educational data mining: A review of the state of the art," *A review of the state of the art. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601-618, 2010.
- [18] C. Romero and S. Ventura , "Educational data mining: A survey from 1995 to 2005," *ELSEVIER*, no. 33, pp. 135-146, 2007.
- [19] S. A. Kumar and M. N. Vijayalakshmi , "Mining of student academic evaluation records in higher education," in *International conference on recent advances in computing and software systems*, 2012.
- [20] E. P. I. Garcia and P. M. Mora, "Model prediction of academic performance for first year students," in *10th Mexican International Conference on Artificial Intelligence*, 2011, pp. 169-174.
- [21] R. S.J.d. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, 2009.
- [22] B. K. Baradwaj and S. Pal, "Mining Educational Data To Analyze Student Performance," *International Journal Of Advanced Computer And Applications*, vol. 2, p. 6, 2011.
- [23] N. S. Shah, "Predicting factors that affect students academic performance by using data mining," *pakistan business review*, january 2012.
- [24] M.B.Sacre,L.J.Shuman,H.Wolfe and C.J.Atman M.Moreno, "Self-Assessed Confidence in EC-2000 Outcomes: A Study of Gender and Ethnicity Differences Across Institutions," in *30th ASEE/IEEE Frontiers in Education Conference*, Kansas City,MO, 2000, pp. 23-28.
- [25] D.Durben and S.A.Junek S.D.Kellogg, "Critical Factors for Success in an Introductory Astronomy Class," in *34th ASEE/IEEE Frontiers in Education Conference*, Savannah,GA, 2004.
- [26] F. Siraj and M.A. Abdoulha , "Uncovering Hidden Information Within University's Student Enrollment Data Using Data Mining," in *3rd asia international conference on modelling and simulation*, Bali, 2009, pp. 413 - 418. (references)
- [27] Yi-Jia Lyu and Yu-Min Wang Ping-Feng Pai, "Analyzing academic achievement of junior high school students by an improved rough set model," *ELSEVIER*, vol. 54, no. 4, pp. 889-900, May 2010. (references)
- [28] O.Soto and E.Tova, "The Use of Competences Assessment to Predict the Performance of First Year Students," in *40th ASEE/IEEE Frontiers in Education Conference*, Washington,DS, 2010. (references)
- [29] N.Buniamini,J.L.A.Manan and N.Hamzah P.M.Arsad, "Proposed Academic Students' Performance Prediction Model: A Malaysian Case Study," in *3rd International Congress on Engineering Education*, 2011.
- [30] Kesavaraj G. and Sukumaran S., "A study on classification techniques in data mining," in *Fourth international conference on computing,communications and networking technologies*, Tiruchengode, 2013, pp. 1-7. (references)
- [31] V. Ramesh, P. Parkavi and P. Yasodha, "Performance Analysis of Data Mining Techniques for Placement Chance Prediction," *International Journal of Scientific & Engineering Research*, vol. 2, no. 8, 2011.
- [32] A.Istanbullu and M.Karahasan H.Guruler, "A new student performance analysing system using knowledge discovery in higher educational databases," *Computers & Education*, vol. 55, no. 1, pp. 247-254, August 2010. (references)
- [33] J. Gamulina, O. Gamulinb, D. Kermekc, "Comparing classification models in the final exam performance prediction," in *37th International Convention on Information and Communication Technology,Electronics and Microelectronics*, Opatija, 2014, pp. 663-668. (references)

CONFERENCE PAPER

978-93-85670-72-5 © 2016 (RTCSIT)

International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)

21st August 2016

Guru Nanak College Budhlada, Punjab India