International Journal of Pure and Applied Mathematics
Volume 117 No. 15 2017, 211-226
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Performance Assessment of Hybrid Multiple Linear Regression Algorithm in Predicting Student Placement Probability

[1]Satish Thatavarti and [2]K. Thammi Reddy
[1]Department of CSE, GITAM University,
Visakhapatnam, India.
satish5813@gmail.com
[2]Department of CSE,
GITAM University,
Visakhapatnam, India.
thammireddy.konala@gitam.edu

Abstract

Data mining, is the discipline of tunneling keen on databases to retrieve knowledge as well as information. It has lately developed new axes of appliances and created a promising discipline, namely, Educational Data Mining[24]. The main educational idea is to craft promising career chances for students in reputed institutes world-wide.In real race, forecasting the students placement chance is a demanding task. Conventional data mining and machine learning methods might not be applied openly to these types of educational facts and problems[14]. The objective of this research is to show how accurate the proposed hybrid algorithm can calculate the placement chances of a student and also to show the performance strategy of the projected algorithm in view of space and time complexity compared to accessible regression analysis algorithms.

Key Words: Educational data mining, multiple regression, ada boosting, placement prediction.

1. Introduction

In Educational Data Mining, Predictive Analytics [12] is extensively used to examine past statistics and spot the future values to generate accurate results. Many formulated methods like Decision Tree, Neural Networks, Naive ayes and K-Nearest Neighbor were practiced on engineering graduates prior performance statistics to spawn the model that can be used to deduce learner's performance and placement chances [5][8]**.**Researchers reported that there are many prediction models are existing with diverse approaches in predicting student performance, however there is no conviction that these predictors are able to determine accurately that a student will be placed[21]. Prediction of outcome will help to channelize labors of students and instructors to get counteractive actions towards the improvement of the undergraduate through the track[22].

For this intention, The statistical model (**Multiple Linear Regression Anaysis**) in conjunction with machine learning (**AdaBoost**) optimization techniques are adopted to build a hybrid multiple linear regression (HMLRA) algorithm to predict the student placement chance accurately and to weigh up the potential of the algorithm.

In **Adaboost**, a booming classifier [25]. fragile learners are collected, "boosted", to perk up group accuracy and construct a "strong" classifier. In Adaboost, sampling is done on the training iteratively, with substitute, to train the fragile learner. The absolute result of Adaboost, is a weighted grouping of all hypotheses, with the weights equivalent to each individual hypothesis' forecast error.It has low implementation difficulty and the introduction of a single tuning parameter, the number of iterations T[10][11].

Multiple linear regression Analysis, a statistical model employed to illustrate data as well as to clarify the dependency among single dependent and multiple independent variables[29].It composes of three phases: To analyze the relationship and directionality of the data, estimate the model and finally to assess the validity and worth of the model[29].

2. Literature Survey

Research has shown that students' performance can be assessed, monitored and can be to predicted which students are likely to place by applying predictor attributes, data mining practices, classification and regression trees (C&RT) on educational data [4][7] [20].

T. Jeevalatha et.al implemented decision tree algorithm to envisage the capable students to be deployed in companies [26] and the researchers concluded that the ID3 performs well compared to other with a precision of 95.33%. SeemaPurohit and NeelamNaik developed a model to categorize the presentation of the students placed [17]. Saurabh Pal and Ajay Kumar Pal [1]gathered the data to analyse and revise the learner's educational

performance necessary to train and place students. The authors concluded that lowest average error , 0.28 can be deduced from Naïve Bayes classifier.

Ajay Shiv Sharma et.al, stated a novel logistic regression model that can be used to generate accurate placement prediction system [2]. OktarianiNurulPratiwi [18] used special classification algorithms to classify the student chance and oncluded that J48 is the best with a precision of nearly 80%. BahenSen et.al, assembled the bulky, attribute dataset to develop the replica for placement prediction[6]. They used Artificial Nueral Networks, vector machine, C5 Decision Tree algorithms to resolve the fact, that the better prediction model will be C5 Decision Tree algorithm with a precision of 95%.

Vikas Chirumamilla et. al, offered the study of bi fold objective [27]. The authors applied naïve Bayes, C4.5 and extraction techniques and simulation results proved C4.5 works well over Naïve Bayes.

Ravi Tiwari et. al, constructed a prediction system by the use of random tree algorithm to improve the placement chance of the students [23]. V. Ramesh et al.[28] applied the data mining techniques to foresee placement chances usingWEKA tool and 5 algorithms. The experimal results show that NaiveBayes Simple, Multilayer Perception, SMO, J48 and REP Tree produced 83.193%, 87.395%, 84.0336%, 84.8739%, 84.8739% accurate results correspondingly.

It can be concluded that Decision tree algorithms were widely adopted. In this study, the stress is on AdaBoost, and Multiple Linear Regression Analysis algorithms to amplify the accuracy.

3. Research Methodology

To ensure the student start the career and move forward in the right direction for better quality living, it is necessary to identify the potential talent by predicting their performance using past performance and knowledge[3][26]. The proposed Placement Prediction System is based on historical data of students performance in academics including attributes like in-semester marks, end-semester marks, practical performance and attendance, training and placement tests etc.

The idea proposed in this paper is to perform analysis on student data using Hybrid Multiple Linear Regression Ananlysis (HMLRA) Model considering number of parameters for the derivation of performance and placement prediction indicators needed for student performance assessment, monitoring and evaluation.

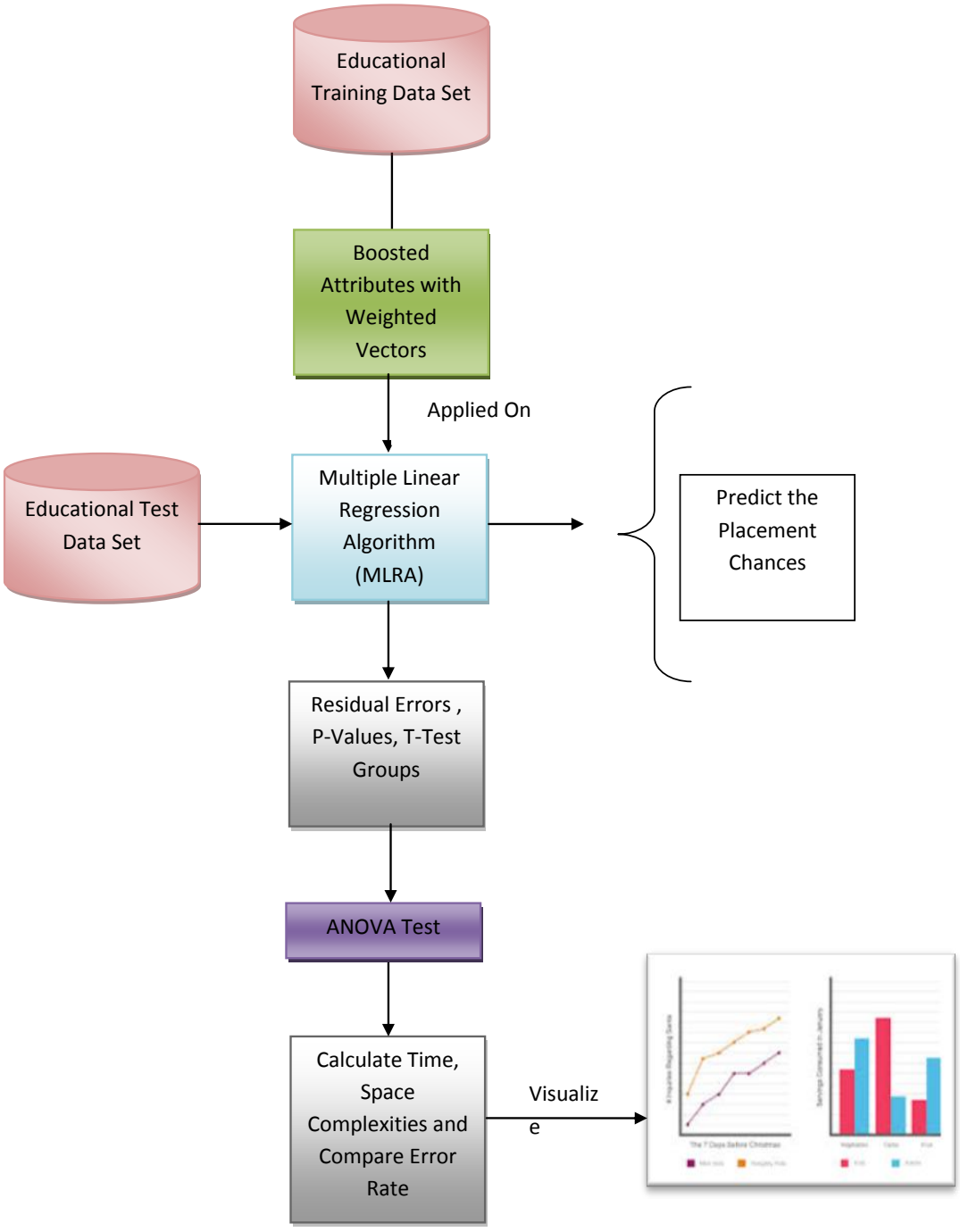


Figure 1: Architecture of the Proposed HMLRA System

HMRLA - Hybrid Multiple Linear Regression Analysis

The aim of this study is to calculate probability of student placement by applying the statistical and machine learning techniques and to analyze the accuracy of the algorithm. Thus, a hybrid algorithm, Multiple Linear Regression Analysis with Ada Boosting has been proposed.

Algorithm

1.Input.

*Sequence of m examples(xi,yi), ..., **(x_i,y_i).** Here labels are y **Y** = {I,.. ., **k) for** classification and **ybelongs to R** for regression problems.*

Integer T specifying number of iterations (machines)

Threshold Φ for demarcating correct and incorrect predictions

2.Initialize:

Machine number or iteration t=1

*Distribution **$D_t(i)=1/m$** forall **i***

for Error rate ε_t avg.loss function $L_T = o$

3. Iterate

*while error rate **ε_t < 0.5** for categorization, or average loss function $L_{T<0.5}$ $r_{t<}T$ for AdaBoosf.RT*

Call Fragile Learner, provided that by distribution D_t

*Get hack a hypothesis **h_t : X -> Y** for categorization*

*Build the regression model: **$y = \beta 0 + \beta 1x1 + \beta 2x2 + . . . + \beta kxk + \varepsilon$** for regression problems*

For classification: Calculate the error rate of h_t :

$$\varepsilon_t = \sum D_t(i) \quad i : h_t(x_i) \neq y_i$$

4.Update distribution Dt as

$$D_{t+1}(i)=Dt(i)/Z_t \; X \{ \beta 0. \beta kxk. \; \textbf{if } |f_t(x_i-y_i/y_i) <= \Phi| \textbf{ ,1 otherwise } \}$$

where Z_t is a normalization factor chosen such that D_{t+1} will be a distribution

set t=t+1

5. Output the final hypothesis

$$f_{fin}(x) = \frac{\sum_t \log\left(\frac{1}{\beta 1}\right) f_t(x)}{\sum_t \log(1/\beta 1)}$$

This HMLRA system is implemented in three stages. In the first stage, Adaboost.RT (Regression Threshold) generates a sample from the training data by means of the sampling weight vector,Dt. The learning algorithm uses this sample to produce a hypothesis that associates the input with output data. The hypothesis is then applied to all the data to produce predictions. The absolute relative error is deliberated for each prediction and weigh against a threshold,. The threshold is used to classify the expected values ares accurate or not. The comparision of threshold alters the regression task to a categorization problem. The sampling weight of inaccuratetly predicted samples is improved for the subsequent iteration[16].

In the second stage, the boosted training data set is applied to Multiple Linear Regression Analysis[13] algorithm. The test data set is used for assessing model performance and predicting values.

In the final stage, Analysis of variance (**ANOVA**) test, a set of statistical models is performed to examine the dissimilarities among group means and their related measures[9].This test helps in calculating space and time complexities of HMLRA and to compare the error rates with Multiple Linear Regression Analysis.

4. Experimental Results

This system is built and implemented using C# language built on .NET framework; Visual Studio 2013- an IDE and SQL Server, a SQL-based relational database management system and WinForms, a graphical GUI technology.

In our study, Adaboost is individually performed on the five student performance datasets i.e., assignment marks, attendance, internal marks, final exam marks and training marks. Each individual resultant training set is applied to model the relationship between attributes of each dataset and placement chance.

The line of regression characterizes the probable placement chance for a specified permutation of the input features[31]. Scatter plot is described by a linear equation of $\mathbf{y=\beta 0 + \beta 1\ x1+ \beta 2x2+.....+ \beta_i x_n}$ for i ranges from 1 to n.

The variation between the regression line and the single data point isthe discrepancy called a residual.

The least squares methods are used to reduce the residual.

$$\sum \mathbf{e_i^2}$$

$$\sum \mathbf{(y_i - \hat{y}_i)^2}$$

$$\sum \mathbf{(y_i - \beta_0 - \beta_1 - x_1 - \beta_2 - x_2 - - - \beta_p - x_n)^2 \Rightarrow min \Rightarrow \hat{y}_i}$$

$$\mathbf{= \beta_0 + \beta_1 - x_1 + \beta_2 - x_2 + \beta_3 - x_3 + - - + \beta_p - x_n}$$

The variance of multiple linear regression σ^2 is projected by

$$s^2 = \frac{\sum e_i^2}{n - p - 1}$$

here p stands for amount of independent variables n is the sample size.

Now the strength and helpfulness of an equation is evaluated.

R^2 is the significant measure to validite the estimated linear line.

R^2 = total variance / explained variance.

To see whether the multiple linear regression model is suitable proficiently a adjusted R^2 is computed by

$$R^2 = R^2 - J(1 - R^2)/N - J - 1$$

here J is the amount of independent variables.

N stands for sample size.

Test of signifance is the final step for the multiple linear regression study. Here two tests are being performed. Firstly, the F-tests of the model checks whether $R^2=0$. Subsequently, numerous t-tests scrutinize the importance of every discrete coefficient and the intercept. The null hypothesis is generated if the coefficient or the intercept is zero[30].

ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the *t*-test to more than two groups. ANOVAs are useful for testing three or more means for statistical significance. ANOVA is more conservative and is suited to a broad variety of realistic problems [15].

Table 1: Boosted Placement Education Marks Data Set to use as Training Set for Student Placement Chance Prediction

Training Data Set					
Questions Attempted	**Questions Correct**	**Questions Incorrect**	**Total Questions**	**Marks Scored**	**Placement**
82	25	57	90	25	0
74	53	21	90	53	0
90	46	44	90	46	1
0	0	0	90	0	1
90	34	56	90	34	1
90	72	18	90	72	0
90	62	28	90	62	1
90	59	31	90	59	1
70	31	39	90	31	1
74	63	11	90	63	0
90	43	47	90	43	1
69	36	33	90	36	1
83	50	33	90	50	0
81	40	41	90	40	1
60	44	16	90	44	1
90	57	33	90	57	1

Table 2: Test Data Set to be Applied to the HMLRA Model for Accurate Student Placement Chance Prediction

Test Data Set				
Questions Attempted	**Questions Correct**	**Questions Incorrect**	**Total Questions**	**Marks Scored**
86	27	60	90	46
87	32	63	90	42
87	25	44	90	48
81	31	0	90	39
85	25	56	90	44
85	33	18	90	42
81	31	28	90	40
80	35	31	90	46
83	26	39	90	48
82	31	11	90	37
84	30	47	90	46
87	32	33	90	35
88	25	33	90	47
89	30	41	90	48
86	25	16	90	35
80	35	33	90	46

Table 3: Shows the Predicted Values Produced by HMLRA Model

SNo	**Placement Chances**	**SNo**	**Placement Chances**	**SNo**	**Placement Chances**	**SNo**	**Placement Chances**	**SNo**	**Placement Chances**	**SNo**	**Placement Chances**
1	0.696	**11**	0.574	**21**	0.643	**31**	0.696	**41**	0.459	**51**	0.67
2	0.604	**12**	0.652	**22**	0.659	**32**	0.657	**42**	0.653	**52**	0.673
3	0.677	**13**	0.666	**23**	0.652	**33**	0.673	**43**	0.675	**53**	0.624
4	0.459	**14**	0.668	**24**	0.659	**34**	0.65	**44**	0.651	**54**	0.644
5	0.705	**15**	0.639	**25**	0.657	**35**	0.666	**45**	0.684	**55**	0.696
6	0.618	**16**	0.723	**26**	0.648	**36**	0.645	**46**	0.625	**56**	0.459
7	0.641	**17**	0.459	**27**	0.652	**37**	0.709	**47**	0.677	**57**	0.661
8	0.648	**18**	0.68	**28**	0.636	**38**	0.661	**48**	0.646	**58**	0.618
9	0.64	**19**	0.689	**29**	0.691	**39**	0.676	**49**	0.668	**59**	0.459
10	0.581	**20**	0.459	**30**	0.664	**40**	0.684	**50**	0.459		

This result is summarized to give the details like number of students to be trained more and the number of students having more chances on placements.

Table 4: Summarized Details of Placement Chances in Percentage

S.No	Placement Chances	Count
1	Less than 50%	6
2	Between 50% to 60%	2
3	Between 60% to 70%	46
4	More than70%	3

Table 5: Performance Evaluation of HMLRA Algorithm

Attributes	Value	Std.Error	t-stat	P-Value
Constant	0.639	0.305	2.095	0.04
QA	-0.004	0.004	-1.053	0.296
QC	0.008	0.004	2.22	0.03
QI	-0.007	0.004	-1.865	0.066
TQ	0.001	0.003	0.401	0.689
MS	0.0003	0.003	0.087	0.931

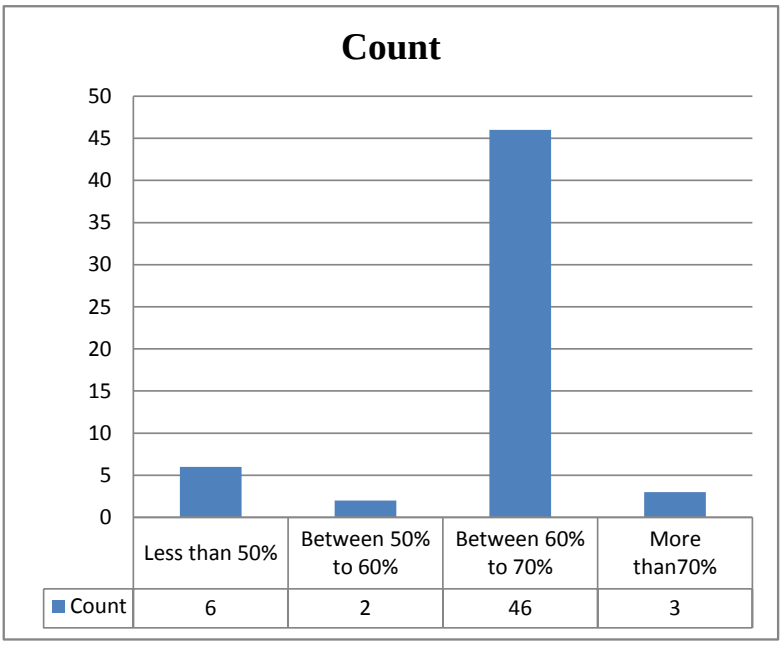


Figure 1: Show the graph of predicted placement chances in percentage

Considering these predicting values, it is easy to find out the strengths and weaknesses of students to improve the placement chances

Table 6: Wealth Informaton about HMLRA Algorithm

Residual Std.Error	R-Square	Adjust R-Square	F-Static
0.488	0.105	0.04	1.612

Table 7: 3.95% Confidence of Interval for Intercept

Interval Lower Bound	95% confidence of Interval for Intercept
-0.011	0.003

Table 8: ANOVA Table for HMLRA Performance Evaluatiot

	Sum of Squares	D.O.F	Var.East
HMLA Regression	1.919	5	0.384
Residuals	16.427	69	0.238
Total	18.347	74	0.248

Table 9: Shows the Comparision of HMLRA with Existing Multiple Linear Regression Analysis (MLRA) in Terms of Space, time Complexities and Error Rate

		HMLRA			MLRA		
Datasets	No of Records	Time Complexity (mSec)	Space Complexity (KB)	Error Rate	Time Complexity (mSec)	Space Complexity (KB)	Error Rate
CSE	195	2458	4903	0.02468	2689	4892	0.033746
ECE	125	3348	3355	0.56890	4525	7550	0.06895
EEE	130	3004	3568	0.65899	6382	7445	0.04789
MECH	168	2733	4640	0.35698	6550	8314	0.24586
CIVIL	65	1887	3401	0.34445	4056	6061	0.07899
PE	98	1823	3752	0.8966	6766	6820	0.024588

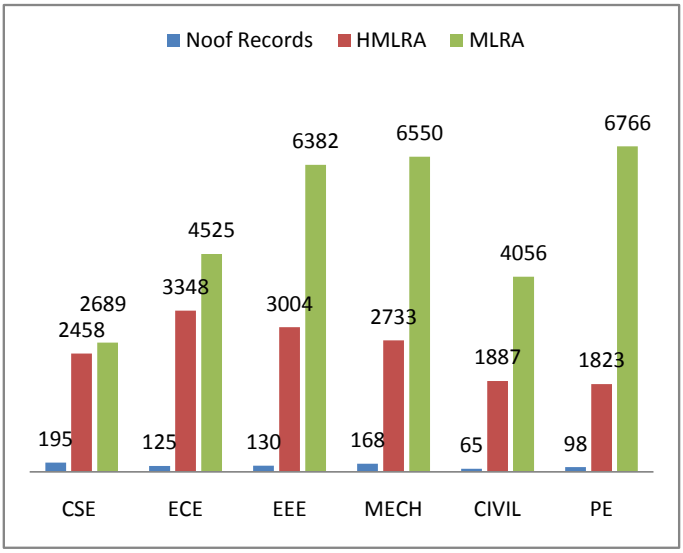
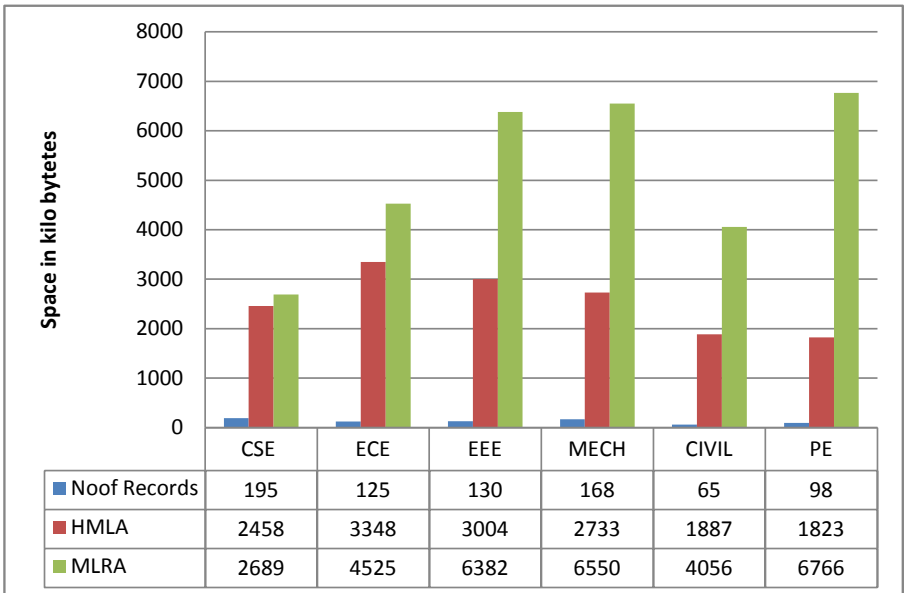


Figure 2: Comparision Graph of HMLRA with MLRA in terms of Time Complexity



	CSE	ECE	EEE	MECH	CIVIL	PE
Noof Records	195	125	130	168	65	98
HMLA	2458	3348	3004	2733	1887	1823
MLRA	2689	4525	6382	6550	4056	6766

Figure 3: Comparision Graph of HMLRA with MLRA in terms of Space Complexity

5. Conclusion

To predict the placement chance of a learner is an immense alarm in higher education systems. The scope of this study is to explore the prediction accuracy of student placement chance and to guesstimate the performance of data mining techniques used in such a situation[19]. Hence, a system HMLRA is proposed in which machine learning algorithm in combination with statistical model is applied. Adaptive Boosting, a machine learning technique helps in boosting the training set to create a strong classifier. Multiple Linear Regression Analysis, a statistical model learns the training data set and builds a hybrid model on which test data is applied. From the experimental outcomes it is proved that HMLRA algorithm is most apt to predict student placement chance. HMLRA also gives accurate prediction of 95% which is moderately superior than other existing algorithms. To conclude, it can be said that Optimum Predictions could be obtained by this HMLRA Model. Moreover, this study is also an attempt to compare the performance and error rate of the proposed algorithm with the on hand machine learning algorithms.

6. Acknowledgement

This paper publication is a part of major research project funded by University Grants Commission (UGC), MHRD, titled: " Usage of Data mining techniques in design and development of Academic Audit System for effective Teaching Learning Process".

References

[1] Pal A.K., Pal S., Classification model of prediction for placement of students, International Journal of Modern Education and Computer Science 5(11) (2013), 49-56.

[2] Sharma A.S., Prince S., Kapoor S., Kumar K., PPS Placement prediction system using logistic regression, In IEEE International Conference on MOOC, Innovation and Technology in Education (MITE) (2014), 337-341.

[3] Amruta R.J., Thomas A., Data Mining for Classification of Students based on their Performance, International Journal for Research in Applied Science & Engineering Technology 5 (4) (2017).

[4] Asif R., Merceron A., Ali S.A., Haider N.G., Analyzing undergraduate students' performance using educational data mining, Computers & Education, (2017).

[5] Aziz A.A., Ismail N.H., Ahmad, F., First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms, In International Conference on Artificial Intelligence and Computer Science (2014).

[6] Şen B., Uçar E., Delen D., Predicting and analyzing secondary education placement-test scores: A data mining approach, Expert Systems with Applications 39(10) (2012), 9468-9476.

[7] Baradwaj B.K., Pal S., Mining educational data to analyze students' performance, (2012).

[8] Baradwaj B.K., Pal S., Mining educational data to analyze students' performance, International Journal of Advanced Computer Science and Applications 2 (6) (2011).

[9] DeBoer J., Stump G.S., Seaton D., Breslow L., Diversity in MOOC students' backgrounds and behaviors in relationship to performance in 6.002 x, In international conference on Proceedings of the sixth learning networks consortium 4 (2013).

[10] Frénay B., Verleysen M., Classification in the presence of label noise: a survey, IEEE transactions on neural networks and learning systems 25(5) (2014), 845-869.

[11] Freund R.M., Grigas P., Mazumder R., A new perspective on boosting in linear regression via subgradient optimization and relatives, (2015).

[12] Agrawal G.L., Gupta H., Optimization of C4. 5 decision tree algorithm for data mining application, International Journal of

Emerging Technology and Advanced Engineering 3(3) (2013), 341-345.

[13] Grégoire G., Multiple linear regression, European Astronomical Society Publications Series 66 (2014), 45-72.

[14] Guo B., Zhang R., Xu G., Shi C., Yang L., Predicting students performance in educational data mining, In IEEE International Symposium on Educational Technology (ISET) (2015), 125-128.

[15] Kim H.Y., Analysis of variance (ANOVA) comparing means of more than two groups, Restorative dentistry & endodontics 39 (1) (2014), 74-77.

[16] Kummer N., Najjaran H., Adaboost. MRT: boosting regression for multivariate estimation, Artificial Intelligence Research 3(4) (2014).

[17] Naik N., Purohit S., Prediction of Final Result and Placement of Students using Classification Algorithm, International Journal of Computer Applications 56(12) (2012).

[18] Naik, N., & Purohit, S. (2012). Prediction of Final Result and Placement of Students using Classification Algorithm. International Journal of Computer Applications, 56(12).

[19] Pal A.K., Pal S., Data Mining Techniques in EDM for Predicting the Performance of Students, International Journal of Computer and Information Technology 2(6) (2013).

[20] Peña-Ayala A., Educational data mining: A survey and a data mining-based analysis of recent works, Expert systems with applications 41(4) (2014), 1432-1462.

[21] Parmar K., Vaghela D., Sharma P., Performance prediction of students using distributed Data mining, In IEEE International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (2015), 1-5.

[22] Qasem A., Al-Radaideh, Ahmad Al Ananbeh, Emad M. Al-Shawakfa, A Classification Model For Predicting The Suitable Study Track For School Students Classification 8 (2) (2011).

[23] Tiwari R., Sharma A.K., A Data Mining Model to Improve Placement, International Journal of Computer Applications, 120(12) (2015).

[24] Baker R.S.J.D., Data Mining for Education, In International Encyclopedia of Education, , B. McGaw, P. Peterson, E.Baker (Eds.), 3e, Oxford, UK: Elsevier 7 (2010), 112-118.

[25] Sprenger M., Schemm S., Oechslin R., Jenkner J., Nowcasting Foehn Wind Events Using the AdaBoost Machine Learning Algorithm, Weather and Forecasting 32(3) (2017), 1079-1099.

[26] Sriram G., Srinivas Y., Thammi Reddy K., A Maximally Specific Hypothesis for Predicting the Employability Requirements in Higher Educational Institutions, International Journal of Modern Computer Science (IJMCS) 3 (1) (2014).

[27] Jeevalatha T., Ananthi N., Kumar D.S., Performance analysis of undergraduate students placement selection using Decision Tree Algorithms, International Journal of Computer Applications 108(15) (2014).

[28] Vikas Chirumamilla, BhagyaSruthi T., Sasidhar Velpula, Indira Sunkara, A Novel approach to predict Student Placement Chance with Decision Tree Induction, International journal of Systems and Technologies Double Blind Peer Reviewed 7 (1) (2014) 78-88.

[29] Ramesh V., Parkavi P., Yasodha P., Performance analysis of data mining techniques for placement chance prediction, International Journal of Scientific & Engineering Research 2(8) (2011).

[30] Yu T., Jo IH. (, March). Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge (2014) 269-270.

[31] RAJESH, M. "A SYSTEMATIC REVIEW OF CLOUD SECURITY CHALLENGES IN HIGHER EDUCATION." The Online Journal of Distance Education and e-Learning 5.4 (2017): 1.

[32] Rajesh, M., and J. M. Gnanasekar. "Protected Routing in Wireless Sensor Networks: A study on Aimed at Circulation." Computer Engineering and Intelligent Systems 6.8: 24-26.

[33] Rajesh, M., and J. M. Gnanasekar. "Congestion control in heterogeneous WANET using FRCC." Journal of Chemical and Pharmaceutical Sciences ISSN 974 (2015): 2115.

[34] Rajesh, M., and J. M. Gnanasekar. "Hop-by-hop Channel-Alert Routing to Congestion Control in Wireless Sensor Networks." Control Theory and Informatics 5.4 (2015): 1-11.

[35] Rajesh, M., and J. M. Gnanasekar. "Multiple-Client Information Administration via Forceful Database Prototype Design (FDPD)." IJRESTS 1.1 (2015): 1-6.

[36] Rajesh, M. "Control Plan transmit to Congestion Control for AdHoc Networks." Universal Journal of Management & Information Technology (UJMIT) 1 (2016): 8-11.

[37] Rajesh, M., and J. M. Gnanasekar. "Consistently neighbor detection for MANET." Communication and Electronics Systems (ICCES), International Conference on. IEEE, 2016.

[38] http://www.statisticssolutions.com/multiple-linear-regression/

[39] http://www.colorado.edu/amath/sites/default/files/attached-files/lesson10 _simple reg_ 0. pdf