# Performance Analysis of Classification Algorithms for Prediction of Student Failure in College using Academic Data Processing

**Miss. Trupti Diwan[1] Prof. Bharati Dixit[2]**
[1,2]M.E. Student [3]Assistant Professor
[1,2]Department of Information Technology
[1,2]MIT College of Engineering Pune, India

*Abstract*— Student data sets give useful information about efficient educational knowledge. Student failure affects education quality. A data mining approach is applied to predict student failure in college at MIT College of engineering Pune, based on internal marks and external marks. Three different classification algorithms, AD Tree, LAD Tree, and ICRM, are employed to build prediction model. This paper summarizes and compares three different classification algorithms. In the end, accuracy of three algorithms is analyzed. The results show that ICRM algorithm has obtained highest TN Rate and GM compared to other algorithms. So the accuracy of prediction is more accurate.

***Key words:*** Educational Data Mining, Classification Algorithms, Student Failure, Decision Tree, WEKA, Prediction

## I. INTRODUCTION

Student failure is one of several important factors affecting education quality. Studies have shown a strong correlation between student failures and attributes affecting them. The Waikato Environment for Knowledge Analysis tool (WEKA) is widely used model in data mining. This paper looks at the student data as inputs to WEKA. The student data enable WEKA to predict student failure.

Instructive data mining (EDM) [1] is a rising interdisciplinary research region that arrangements with the advancement of strategies to investigate information beginning in an instructive context. EDM uses computational ways to deal with break down educational information so as to study instructive inquiries.

Instructive information mining (EDM) is a field that endeavors measurable, machine-learning and information mining (DM) calculations over the distinctive sorts of instructive information. Its principle goal is to dissect these sorts of information so as to determine instructive exploration issues [1]. EDM is concerned with creating strategies to investigate the novel sorts of information in instructive settings and, utilizing these strategies, to better get it understudies and the settings in which they learn. On one hand, the increment in both instrumental instructive programming also as state databases of understudy's data have made vast archives of information reflecting how understudies learn. On the other hand, the utilization of Internet in training has made another setting known as e-learning or online instruction in which a lot of data about teaching–learning collaboration are unendingly created and pervasively accessible. All this data gives a gold mine of instructive information. EDM tries to utilize these information archives to better get it learners and learning, and to create computational methodologies that join information and hypothesis to change practice to advantage learner.

Late years have demonstrated a developing intrigue and concern in numerous nations about issue of school disappointment and the determination of its principle contributing elements [2]. The considerable arrangement of examination [2] has been done on distinguishing the components that influence the low execution of understudies (school disappointment and dropout) at diverse instructive levels (essential, auxiliary furthermore, higher) utilizing the expansive measure of data that current PCs can store in databases. All these information are a "gold mine" of significant data about understudies. Distinguish and find helpful data covered up in substantial databases is a troublesome assignment. An extremely encouraging answer for accomplish this objective is the utilization of information disclosure in databases methods or information mining in instruction, called instructive information mining, EDM

This paper mainly focuses on student failure prediction in final exam. After the prediction of student's failure we check it with actual university results of students. By predicting this we can reduce the failure of students by providing them extra classes or special coaching.

## II. RELATED WORK

In this paper projected data processing procedure [2] to calculate college dropout and disappointment. Here utilize real knowledge on 670 college students from Zacatecas, Mexico and apply white box classification methods for e.g. call tree and induction rules. Here it's summary of the state of expertise with reference to EDM and surveys absolutely the perform of this kind of date. In every study that's classified by the type of information and DM strategies applied, in addition by the kind of tutorial occupation that they resolve.

Understudies casual discussions on online networking (e.g., Twitter, Facebook) shed light into their instructive encounters— sentiments, emotions, and worries about the learning procedure. Information from such uninstrumented situations can give significant information to advise understudy learning. Breaking down such information, then again, can be testing. The multifaceted nature of understudies encounters reflected from online networking substance obliges human translation. Be that as it may, the developing size of information requests programmed information investigation methods. In this paper [3], author has built up a work process to incorporate both subjective examination and extensive scale information mining systems. Author has concentrated on designing understudies Twitter presents on comprehend issues and issues in their instructive encounters. Author initially led a subjective investigation on tests taken from around 25,000 tweets identified with designing understudies school life. Author discovered designing understudies experience issues, for example, overwhelming study burden, absence of social

 

engagement, and lack of sleep. Taking into account these outcomes, we actualized a multi-mark arrangement calculation to group tweets mirroring understudies issues. We then utilized the calculation to prepare an indicator of understudy issues from around 35,000 tweets gushed at the geo-area of Purdue University. This work, surprisingly, introduces a procedure and results that show how casual online networking information can give bits of knowledge into understudies encounter.

Data mining refers to the technique of getting hidden, antecedently unknown and presumably important information from whopping quantity of knowledge. Data processing uses a mix of a huge knowledge domain, advanced analytical skills, and domain information to unveil hidden trends and patterns which might be applied in virtually any sector starting from business to drugs, then to Engineering. notwithstanding academic institutes will use data processing to get valuable data from their databases referred to as academic data processing (EDM). Academic data processing needs transformation of existing or innovation of recent approaches derived from statistics, machine learning, psychological science, scientific computing etc. In this paper [4] current system is intended to justify that varied data processing techniques which incorporates classification, are often employed in academic databases to counsel career choices for the high school students and conjointly to predict the doubtless violent behaviour among the scholars by as well as additional parameters apart from tutorial details. RapidMiner has been used as data processing tool.

## III. PROPOSED METHODOLOGY

### *A. System Architecture*

System architecture is shown in Figure1. The methodology starts from problem definition. 600 students' academic records are gathered from MIT College of Engineering Pune.

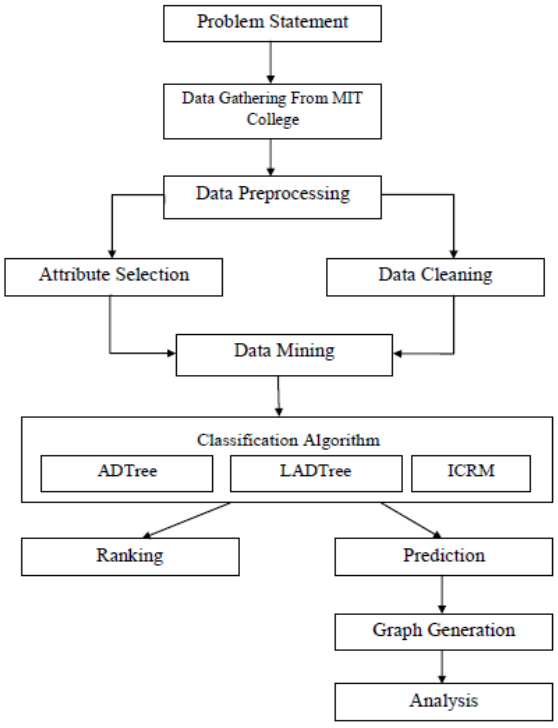


Fig. 1: System Architecture

These records are of second years engineering students. For removing unwanted data and for adding most occurred values if any record is empty preprocessing is used. Based on the rule student dropout and failure is being predicted. Attributes used for prediction are listed below in the TABLE I.

| Attributes | Possible values |
|---|---|
| Assignment I Grades | If Marks 15-20 =A,<br>If marks 9-14=B,<br>If Marks 1-8=C |
| Assignment II Grades | If Marks 15-20 =A,<br>If marks 9-14=B,<br>If Marks 1-8=C |
| Unit Test I Marks | Between 0 to 30 |
| Unit Test II Marks | Between 0 to 30 |
| Attendance | 0 to 100 % |
| Self_Motivation | If attendance > 50<br>True<br>Else<br>False |

Table 1: Attributes & Possible Values

Attributes which are more effective for prediction are selected. After this data mining technique such as classification is used. In classification algorithms ADTree, LADTree, and ICRM are implemented using WEKA tool. These algorithms are used to predict student failure. Parameters of these algorithms are compared in order to check improved performance. Students ranking is based on average percentage calculated and by sorting average percentage in descending order. In the end different graphs are generated and analysis of algorithms is done on different parameters.

### *B. Mathematical Model*

Let, System S is painted as: S =

*1) Data Gathering:*

Let, A may be a set of scholar's dataset A=

Where,

a1, a2… area unit the scholar's information gathered.

*2) Attribute Selection:*

Let B may be a set attribute choice B =

Where,

f1, f2, f3, f4, f5 area unit the chosen attributes.

*3) Classifier and Genetic Algorithm:*

Let, C may be a classifier C =

Where,

e1, e2 area unit the classification method.

*4) Ranking:*

Let, D may be a ranking method,

D =

Where,

d1, d2,… area unit variety marks obtained by students.

### *C. Algorithm*

In this paper three classification algorithms are implemented to check the failure of the students in college. ADTree, LADTree (Least Absolute Deviation), and ICRM (Interpretable Classification Rule Mining) algorithms are used. Workflow of ICRM algorithm is as below:

 

*1) Step1: arbitrarily generated rules are created i.e. known as initial population.*
*2) Step2: A best rule for the various categories is found exploitation algorithmic program*
*3) Step3: Context free grammer is employed*

(S) → (cmp) | (cmp) AND (S)
(cmp) → (op_ca) (variable)(value)
(op_cat)→ >|<|>=|<=
(Variable)→Any valid attribute in dataset
(Value)→Any valid price

*4) Step4: By exploitation genetic operators rules are made. It's accustomed improve accuracy of rule.*
*5) Step5: Sensitivity (Se) and specificity (Sp) are accustomed calculate valley of the fitness operate.*

$$\text{Fitness} = Se \cdot Sp \quad (1)$$

## IV. COMPUTATIONAL RESULTS AND DISCUSSION

Generally performance of classification algorithms is measured in terms of confusion matrix. In this case following four measures are used:

Accuracy (Acc) is that the overall accuracy rate or classification accuracy and is calculated as

$$Acc = (TP+TN)/(TP+TN+FP+FN) \quad (2)$$

True Positive rate (TP rate), additionally referred to as sensitivity (Se) or recall, is that the proportion of actual positives that are foreseen to be positive and is calculated as

$$TPrate = TP/(TP+FN) \quad (3)$$

True Negative rate (TN rate), or specificity (Sp), is that the proportion of actual negatives that are foreseen to be negative and is calculated as

$$TNrate = TN/(TN+FP) \quad (4)$$

Mean (GM) indicates the balance between classification performances within the majority and minority classes; that's, geometric mean may be a live of the central tendency used with unbalanced datasets and is calculated as

$$\text{Geometric mean} = \sqrt{(TPrate*TNrate)} \quad (5)$$

Table 2 shows performance analysis on 2012-13 dataset. Performance of ICRM & LADTree algorithms are improved in terms of TN Rate & GM compared to ADTree algorithm. ICRM has highest TN rate & GM as 52.94% & 66.91% respectively.

| | Accuracy | TP Rate | TN Rate | GM |
|---|---|---|---|---|
| ADTree | 89.65 | 95.16 | 29.41 | 52.90 |
| LADTree | 91.13 | 95.16 | 47.05 | 64.46 |
| ICRM | 76.35 | 78.49 | 52.94 | 66.91 |

Table 2: Performance Analysis on 2012-13 Dataset

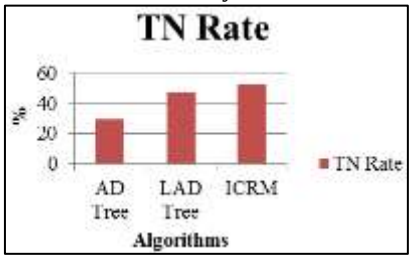


Fig. 2: Graph of TN v/s Algorithms on 2012-13 Dataset

Figure 2 shows graph of TN v/s ADTree, LADTree, & ICRM algorithms on 2012-13 dataset. In this graph ICRM algorithm has highest TN rate (52.94%). TN rate shows students which are predicted as fail and actually failed.

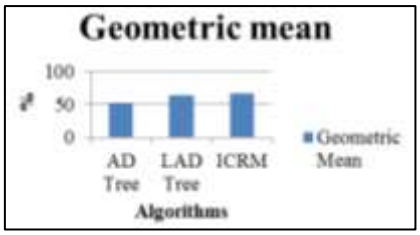


Fig. 3: Graph of GM v/s Algorithms on 2012-13 Dataset

Figure 3 shows graph of GM v/s ADTree, LADTree, & ICRM algorithms on 2012-13 dataset. In this graph ICRM algorithm has highest GM value (66.91%). GM shows balance between majority and minority classes.

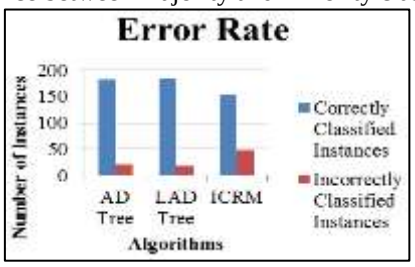


Fig. 4: Error Rate

Figure 4 shows graph of Error rate. That is no of instances v/s correctly and incorrectly classified instances. ICRM algorithm has minimum error rate compared to ADTree algorithm.

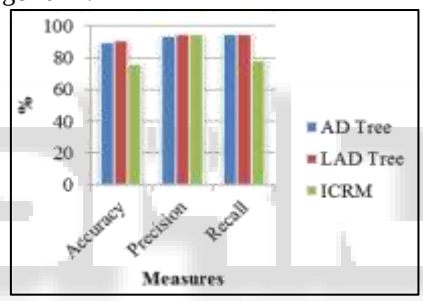


Fig. 5: Comparison of Performance Parameters on 2012-13 Dataset

Figure 5 shows comparison of performance parameters on different algorithms. Accuracy, precision, & recall of three algorithms is compared. ICRM has lowest accuracy value whereas it has highest precision value compared to other algorithms.

TABLE 3 shows performance analysis on 2013-14 dataset. Performance of ICRM & LADTree algorithms are improved in terms of TN Rate & GM compared to ADTree algorithm. ICRM has highest TN rate & GM as 90.69% & 93.85% respectively.

| | Accuracy | TP Rate | TN Rate | GM |
|---|---|---|---|---|
| ADTree | 97.51 | 100 | 54.54 | 73.85 |
| LADTree | 97.51 | 100 | 54.54 | 73.85 |
| ICRM | 96.51 | 96.84 | 90.69 | 93.85 |

Table 3: Performance Analysis on 2013-14 Dataset

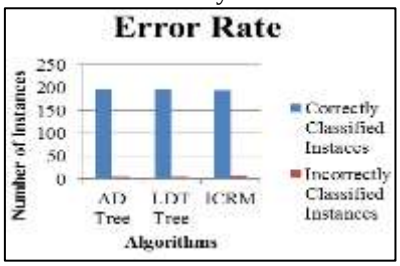


Fig. 6: Error Rate

 

Figure 6 shows graph of Error rate. That is no of instances v/s correctly and incorrectly classified instances. ICRM algorithm has minimum error rate compared to ADTree algorithm.

TABLE 4 shows performance analysis on 2014-15 dataset. Performance of ICRM & LADTree algorithms are improved in terms of TN Rate & GM compared to ADTree algorithm. ICRM has highest TN rate & GM as 73.52% & 76.56% respectively.

| | Accuracy | TP Rate | TN Rate | GM |
|---|---|---|---|---|
| ADTree | 81.34 | 100 | 58.52 | 72.35 |
| LADTree | 85.82 | 100 | 73.52 | 81.34 |
| ICRM | 77.61 | 96.84 | 73.52 | 76.56 |

Table 4: Performance Analysis On 2014-15 Dataset

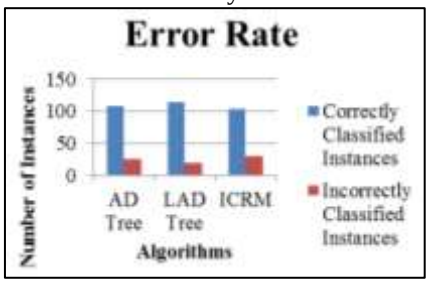


Fig. 7: Error Rate

Figure 7 shows graph of Error rate. That is no of instances v/s correctly and incorrectly classified instances. ICRM algorithm has minimum error rate compared to ADTree algorithm.

## V. CONCLUSION

In this paper student failure prediction model has been built using student's internal and external data. The student's data is obtained from MIT College of Engineering Pune, and data have been collected over the period of three years. Collected data has been preprocessed to remove unwanted information about student.

Among the three classification algorithms tested in this paper, the ICRM and LADTree algorithms has performed the best. In all the cases we can see that accuracy of TN Rate and GM is increased than ADTree algorithm. So the prediction of student failure is done correctly.

The proposed methodology has demonstrated high TN Rate and GM for ICRM algorithm. It has provided student failure prediction and improved education quality. In future research, data from other colleges will be collected for further validation and improvement of proposed approach.

## REFERENCES

[1] Crist´obal Romero, Member, IEEE, and Sebasti´an Ventura, Senior Member, IEEE, "Educational Data Mining: A Review of the State of the Art," IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 40, NO. 6, NOVEMBER 2010

[2] Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto, "Predicting School Failure and Dropout by Using Data Mining Techniques," IEEE JOURNAL OF LATIN-AMERICAN LEARNING TECHNOLOGIES, VOL. 8, NO. 1, FEBRUARY 2013

[3] Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, and Krishna Madhavan, "Mining Social Media Data for Understanding Students' Learning Experiences," IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 7, NO. 3, JULY-SEPTEMBER 2014

[4] Elakia, Gayathri, Aarthi, Naren, "Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students," International Journal of Computer Science and Information Technologies, Vol. 5, NO. 3, 2014, 4649-4652

[5] Kinjal Jakhariya, Shantanu Santoki, "Survey on Prediction the Performance of Students Using Educational Data," Journal of Emerging Technologies and Innovative Research (JETIR), March 2015, Volume 2, Issue 3

[6] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao "PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS" IJDKP, Vol.3, No.5, September 2013.

[7] Mr. M. N. Quadri1,Dr. N.V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques," GJCST, Vol. 10 Issue 2 (Ver 1.0), April 2010.

[8] M.Sindhuja et al. "Prediction and Analysis of students Behaviour using BARC Algorithm," IJCSE, Vol. 5 No. 06 Jun 2013.

[9] Academic patterns using data mining techniques," IJCSE,Vol. 4 No. 06 June 2012.

[10] Edin Osmanbegović , Mirza Suljić, "Data Mining Approach For Predicting Student Performance," Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012 .

[11] Smitha Harikumar, "A Study on Educational Data Mining," International Journal of Computer Trends and Technology (IJCTT) – volume 8 number 2– Feb 2014.

[12] Farhana Sarker, Thanassis Tiropanis, Hugh C Davis. "Linked Data, Data Mining and External Open Data for Better Prediction of at-risk Students", 978-1-4799-6773-5/14/$31.00 ©2014 IEEE.

[13] Alana M. de, Morais and Joseana M. F. R. Araújo, Evandro B. Costa, "Monitoring Student Performance Using Data Clustering and Predictive Modelling", 978-1-4799-3922-0/14/$31.00 ©2014 IEEE.

[14] Oktariani Nurul Pratiwi, "Predicting Student Placement Class using Data Mining", Teaching, Assessment and Learning for Engineering (TALE), 2013 IEEE International Conference on 26-29 Aug. 2013.

[15] Suman Khatwani, Dr. Arti Arya, "A Novel Framework for Envisaging a Learner's Performance using Decision Trees and Genetic Algorithm", 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan.04- 06,2013, Coimbatore, INDIA.

[16] Kin Fun Li, David Rusk and Fred Song, "Predicting Student Academic Performance", 2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems.

[17] Mishra, T., Kumar, D. Gupta, S.," Mining Students' Data for Prediction Performance", Advanced

 

Computing & Communication Technologies (ACCT), 2014 Fourth International Conference on 8-9 Feb. 2014.
[18] B.K.Bhardwaj and S.Paul, "Mining Educational Data to Analyze Students Performance", International Journal Advanced Computer Science and application Vol. 2 No. 6, 2011.
[19] Mohammed M. Abu Tair, Alaa M. El-Halees, "Mining Educational Data to Improve Students' Performance: A Case Study", International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012.
[20] Miss. Trupti Diwan, Prof. Bharati Dixit, "Survey Report on: EDM for Prediction of Academic Trends & Patterns," International Journal of Science and Research Volume 3 Issue 12, December 2014
[21] Miss. Trupti Diwan, Prof. Bharati Dixit, "Analysis of Classification Algorithms for Prediction of Student Failure Using EDM," iPGCON-2015 Fourth Post Graduate Conference AVCOE, Sangamner SPPU, Pune



 