International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228417

Stream Processing for Performance Analysis of Identifying Dropout Students utilizing Different Decision Tree Based Methods

Sumita Guddin*[1], Dr. R. N. Yadawad[2], Dr. U.P. Kulkarni[3]
*[1]PG Student, Department of CSE/SDMCET/Dharwad, Karnataka, India
[2,3]Professor Department of CSE/SDMCET/Dharwad, Karnataka, India

Article Info

Publication Issue :
Volume 8, Issue 4
July-August-2022

Page Number : 143-149

Article History
Accepted: 05 July 2022
Published: 22 July 2022

ABSTRACT

A person's ability to lead a stable, affluent life is made possible through education. In the same way, a country's development may be influenced by the proportion of its population with a higher level of education. This number does, however, decline because of early schooling dropouts. Furthermore, a nation's resources are diminished if a student cannot continue because of a dropout. Although the number of dropouts is constantly falling, it is still very challenging for educational institutions to identify these individuals. An educational institution's first priority is to improve student Performance; therefore, it makes sure that every student graduates on time. Nevertheless, a significant barrier that has a negative effect on this goal is student dropout. Understanding the causes of dropouts is necessary to finding a solution. The causes differ from one student to another; some are connected to the student's workload and mental fortitude. Various ways using Decision Tree (DT) methodologies have been suggested and studied in this study.

Keywords: Important attributes, Decision tree, various decision tree-based approaches, dropouts.

I. INTRODUCTION

We learn, acquire knowledge, and hone our talents with the aid of education. It is crucial for the development of both the world and the country. Our minds and personalities are fundamentally altered by education. It helps us cultivate a positive mindset. Modeling, comprehending, and predicting student performance and academic advancement have garnered a lot of attention lately [1]. Every nation's educational system has a significant impact on its development. This is why many educational institutions emphasize timely graduation. The ratio of graduates and enhancing student performance in educational institutions are both major issues. The likelihood that a graduate will finish their degree on time might depend on a variety of conditions, including financial constraints, a causal attitude, unforeseen life changes, and others. It takes time to identify the kids by observing their behavior. There is a lot of labors involved. Data mining methods can be used in this situation to find dropout students.

A method called data mining enables us to extract hidden information from large amounts of data [2]. Both descriptive and predictive nature are included in

data mining. Methods of supervised learning are typically predictive. However, unsupervised learning is more illustrative [3]. The supervised learning approach delivers input instances together with their class labels for training purposes. By analyzing the supplied data, we can predict what the test data will show. This work has designed, put into practice, and evaluated methods for identifying dropout students based on Random Forest (RF), Logistic Model Tree (LMT), Decision Tree (CART), and AdaBoost Decision Tree (ABT).

To manage the data streams in real time, Apache Kafka, an open-source stream platform, has been used.

The remainder of the paper is structured as follows: II will focus on literature reviews. How to assess the significance of an attribute is covered in Section III. Many tree-based methods are presented in section IV. Section V has a succinct description of the synthetic dataset. In section VI, evaluation metrics were discussed. The tree examined in section VII serves as a representation of how well tree-based strategies perform. Our work's conclusion is covered in Section VIII.

II. LITERATURE SURVEY

The nation slips backward because of student dropouts, and the educational system's resources are depleted. Determining which pupils have dropped out is a difficult undertaking. Researchers have experimented with a number of categorization methods to try and pinpoint dropout kids for this. In today's society, digital data is the de facto money, and it is used by all businesses that use IT. Due to the global increase in data, the ability to analyze huge data volumes, or "big data," has emerged as a competitive advantage. Big data is supported by new waves of productivity, growth, and innovation. Insights into a variety of variables, including student performance and various learning styles, that may affect memory in STEM-related subjects are provided by the research [4–9]. [4] The study, "Student Attraction, Persistence, and Retention in STEM Programs: Successes and Continuing Challenges," Numerous unfavorable assumptions and preconceptions about the nature of the IT profession exist, particularly regarding IT studies. Young people are commonly deterred by these beliefs from enrolling in IT programs and careers, and some even give up on IT (or majoring in IT) [5].

A new paper's authors [6] aim to assess the efficacy of various categorization methods (Naive Bayes, Neural Network, Support Vector Machine, Decision Tree, and Random Forest) in terms of analysis student. Rules for classification have also been developed. A full analysis of student achievement over several years is provided in [7]. Several classification methods, including Decision Tree, Decision Table, Logistic Regression, and Naive Bayes, have been used to identify students who don't meet the required levels. In a paper [8] where they investigate multiple tree-based techniques, the authors show that each tree performs better for each situation. They have conclude that the choice of parameter is crucial for knowledge discovery. The authors of paper [9] compared the algorithms of REP Tree and Decision Tree. They illustrated how the trimming algorithm helps to carry out precise data analysis. Additionally, they focused on how REP Tree impacts modifications to accuracy and complexity. The effectiveness of methods such as the Random Forest algorithm, Logistic Model Tree (LMT), and Classification and Regression Tree has been compared by the authors in [10]. (CART). Random Forest performs better in terms of prediction than CART and LMT.

III. AIMANCE OF ATTRIBUTES

For each dataset, there will be a number of independent attributes and one dependent attribute. One dependent feature and 9 independent traits are

present in the student dataset we used for this investigation. These dependent attributes reveal whether the student discontinued the course or not. There is a chance that not all 9 independent attributes will influence a dependent attribute's value equally. Some independent attributes affect the value of the dependent property more so than other independent attributes. Only 4 of the 9 features that were employed in our work are essential for figuring out how much the dependent property is worth. Equation 1 can be used to compute the information gain.

$$IG(t) = -\sum_{i=1}^{|c|} p(c_i) \log P(c_i) + P(t)P(c_i/t) \log P(c_i/t) + P(1-t) P(c_i/(1-t)) \log P(c_i/(1-t)) \quad \dots (1)$$

Hence "c_i" refers to class. P(t) is the likelihood that the given document is true, P(1-t) is the likelihood that the given document is false, $P(c_i/t)$ is the conditional probability given that the given document is true, and $P(c_i/(1-t))$ is the conditional probability given that the given document is false. The i^{th} class's probability is P(ci). Using Equation, one can calculate the features' Information Gain (IG) (1). The gain value can be used to rank the qualities in order of importance. Using these crucial elements, the dimensionality of the datasets can be reduced. To avoid high dimensional data complicating categorization problems, this is done. A heatmap has numbers that represent various shades of the same colors for each value that will be plotted. Colors on a chart often indicate maximum numbers than lighter ones. Additionally, a very different colors can be used for a very different value.

IV. APPROACHES

A. Random Forest (RF)

The supervised machine learning method known as random forest is a well-liked approach for problems with classification and regression. It uses a variety of samples to generate decision trees, using most of them for categorization and the average of them for regression. One of the most important features of the Random Forest Algorithm is its capacity to handle data sets containing both continuous variables, as in regression, and categorical variables, as in classification [8]. It generates superior outcomes for categorization problems.

B. Logistic Model Tree (LMT)

Incorporates the tree induction and logistic regression [16].

The standard classification tree must first be constructed using the C4.5 approach to generate an LMT. In the logistic variation, the splitting criterion is information gain. Splits might be binary or several ways. Each node creates an LR model using the Logit Boost algorithm. Using the CART algorithm, the tree is trimmed. LMTs outperform other classifiers while being simple to understand. The time needed to construct them, however, is the biggest disadvantage.

C. Decision Tree (CART)

Decision tree algorithms are a subset of supervised learning algorithms. Given that it can be used for both classification and regression tasks, the decision tree methodology stands out among supervised learning techniques. the document, etc. Based on the comparison, we proceed to the next node by following the branch that leads to the value of that value.

D. AdaBoost Decision Tree. (ABT)

AdaBoost works best to improve a weak learner's performance. AdaBoost is a straightforward and organic extension of AdaBoost that Freund and Schapire presented for k>2 classes. M1. Every time a boosting approach is used, a weight is kept. When the weight is bigger, the classifier's impact over the learner is greater. By merging the entire set at the

end, it improves the performance of weak learners [17].

V. DATABASE DESCRIPTION

It has been generated a dataset called "Student Performance Analysis." 9 attributes are included in this dataset, including Name, Address, Semester, Current Semester, and Total backlogs, overall grade totals, grades earned, number of working days, overall class attendance, drop-out status, and other statistics. Instances in the dataset total 106. The dataset is used to assess the efficacy of different tree-
based strategies (Random Forest, LMT, Decision Tree (CART), and AdaBoost decision
tree) for identifying dropout students. In a dataset, a pair plot displays pairwise relationships. To put the methods into practice, WEKA was also employed at Co lab.

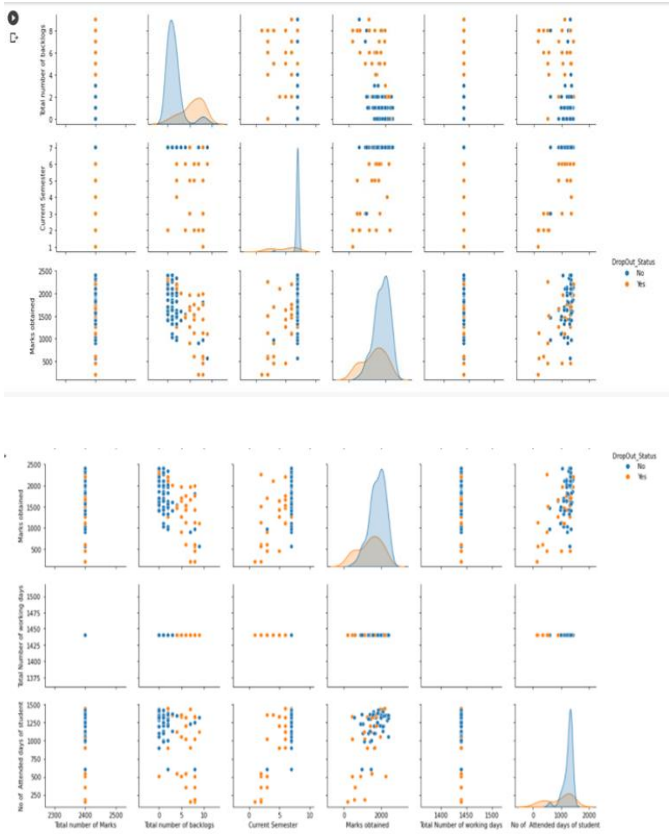


Figure 1 : Student performance analysis dataset in form of pair plot

VI. ANALYSIS METHODS

To assess the success of our work, a variety of metrics have been considered, such as correctness, high accuracy [11], recalls [11], F1-score [12], ROC Region [12], PRC Region [13], MCC [13], and Root Mean Squared Error [13].

A. Precision, Recall and F1 measure

$$\text{Precision} = TP / TP + FP \quad \ldots (2)$$

$$\text{Recall} = TP / TP + FN \quad \ldots (3)$$

$$F1measure = 2 * Precision * Recall / Precision + Recall \quad \ldots (4)$$

Where TP = True Positive is the number of students who graduate on time and FP = False Positive denotes the number of dropout students who are wrongly classified.

False Negative, or FN, refers to an incorrect estimate of the number of students who can continue their education.

B. ROC Region

A ROC curve illustrates the true positive rate as a function of the false positive rate. How accurate the test is will rely on how well it can distinguish between the groups. The optimum test has a ROC Area of 1, and the worst case is when it is less than 0.5.

C. PRC Region

Precision is represented as a function of recall in the PRC curve. When a person is simply concerned with the behaviors of the classifier in a single class, PRC is more helpful. When analyzing binary classifications on unbalanced
data, it is more illuminating than ROC.

D. MCC Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient is a metric used to assess the precision of binary categorization (MCC). The correlation coefficient considers both true positives and false negatives. It gives back a value in the range of -1 to +1. A perfect prediction is indicated by a score of 1, a poor prediction is represented by a score of 0, and a total difference between both the prediction as well as the observed is represented by a score of 1.

$$MCC = T P * T N - F P * F N / \sqrt{(T P + FP)(TP + F N)(T N + F P)(T N + F N)} \ldots (5)$$

E. Root Mean Squared Error (RMSE)

Measures the deviation in between projected and select specific. It could be expressed as

$$RMSE = \sqrt{\sum_{t=1}^{n}(f_i - y_i)^2 / n} \ldots. (6)$$

VII. Performance Evaluation

First, information gain (IG) has been estimated. Most crucial element is regarded as the one that provides the most information. Features have been ordered by their gain value once the information gain has been calculated. The highest information gain feature is taken into consideration first, followed by the other features. The dataset has been used to evaluate the effectiveness of several tree-based algorithms. Fig. 2 shows how our work has generally progressed.

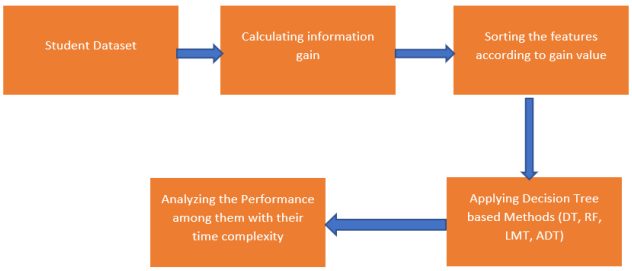
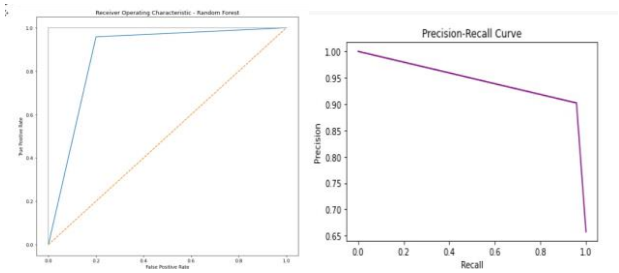


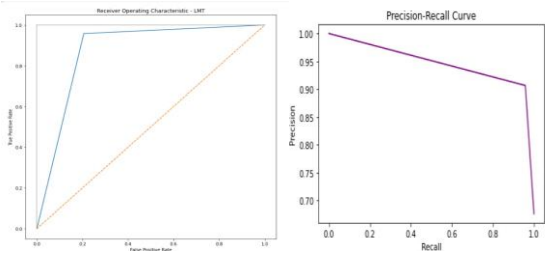
Figure 2 : Flow of Work

The findings of several decision trees, such as Random Forest (RF), Logistic Model Tree (LMT), Decision Tree (CART), and AdaBoost Decision Tree (ABT), are implemented in accordance with the workflow with taking time complexity into account. Such as CPU times, wall time.



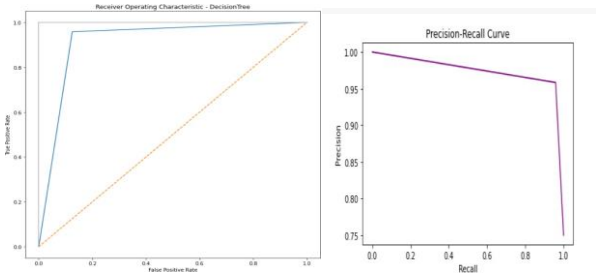
i) ii)

Figure 3: i) Receiver Operating Characteristic Random Forest, ii) precision _ recall _curve



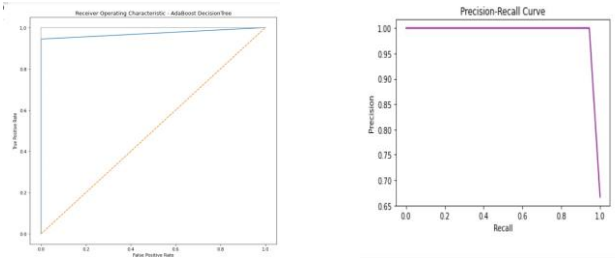
i) ii)

Figure 4: i) Receiver Operating Characteristic Logistic Model Tree, ii) precision _ recall _curve



i) ii)

Figure 5: i) Receiver Operating Characteristic Decision Tree, ii) precision _ recall _curve



i) ii)

Figure 6: i) Receiver Operating Characteristic AdaBoost Decision Tree, ii) precision _ recall _curve

Various decision tree-based approaches for identifying dropout students are compared in the table below.

		DT(CART)	Random Forest	Logistic Model Tree	AdaBoost Decision Tree
Roc curve accuracy value		0.9069	0.87916	0.8759	0.9722
RMSE		0.30618	0.30966	0.3086	0.19245
MCC		0.7984	0.7842	0.7787	0.92195
Precision	0 1	0.83 0.95	0.91 0.90	0.90 0.91	0.90 1.00
Recall	0 1	0.91 0.90	0.80 0.96	0.79 0.96	1.00 0.94
F1-score	0 1	0.87 0.93	0.85 0.93	0.84 0.93	0.95 0.97

Figure 7 : Evaluation of several decision tree-based techniques.

Apply sophisticated processing: TensorFlow, NumPy, SciPy, or Matplotlib are just a few examples of open Source projects that can be used to run machine learning models on streaming data. Installing the necessary packages will enable you to construct a topic, train it, test it, write to it, read from it, and present a summary of the dataset model.

VIII. CONCLUSION

Various tree-based approaches for identifying dropout students are compared in this research. model for summarizing the student dataset while using Kafka to stream data. Calculated significant features are used to identify dropout students. The most crucial elements are regarded as having the highest gain value. Numerous evaluation criteria are applied to gauge the strategies' effectiveness. AdaBoost Decision Tree, Random Forest, and LMT are examples of decision trees. XG Boost, a modern tree-based approach, can be compared to in the future.

IX. REFERENCES

[1]. A. Bowers, and R. Sprott, "Why tenth graders fail to finish high school: a dropout typology latent class analysis," Journal of Education for Students Placed at Risk (JESPAR), vol. 17, no. 3, pp. 129-148, 2012.

[2]. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining: Towards a unifying framework," in the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), Aug. 1996, pp. 82-88.

[3]. C. J. Carmona, P. Gonzales, M. J. Jesus, and F. Herrera, "NMEEF-SD: Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy Rules in Subgroup Discovery," in IEEE international conference on fuzzy systems, pp. 1706-1711, 2010.

[4]. Afterschool, A. (2011). STEM learning in afterschool: An analysis of impact and outcomes. Retrieved from http://www.afterschoolalliance.org/STEM-Afterschool-Outcomes.pdf

[5]. Z. Kovacic, "Early prediction of student success: Mining students' enrolment data." Proceedings of Informing Science & IT Education Conference, 2010.

[6]. Zwedin, S. 2014. Computing Degrees and Enrollment Trends: From the 2012-2014 CRA Talbee Survey. Computing Research Association, Washington D.C.

[7]. Xenos, M., Pierrakeas, C., & Pintelas, P. (2002). A survey on student dropout rates and dropout causes concerning the students in the Course of Informatics of the Hellenic Open University. Computers & Education, 39(4), 361-377.

[8]. Y. Zhao, and Y. Zhang, "Comparison of Decision Tree Methods for finding active objects," National Astronomical Observation, vol. 41, pp. 1955-1959, 2008.

[9]. W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, "A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms," In IEEE International Conference on Control System, Computing and Engineering, 23 - 25 Nov. 2012.

[10]. W. Chen, X. Xie, J. Wang, B. Pradhan, H. Hong, D. T. Bui, Z. Duan, and J. Ma, "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility," Catena 151, pp. 147-160, 2017

[11]. T. Fawcett, "An introduction to ROC analysis," Pattern recognition letters, vol. 27, no. 8, pp. 861-874, 2006.

[12]. A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern recognition, vol. 30, no. 7, pp. 1145-1159, 1997.

[13]. K. H. Walse, R. V. Dharaskar, and V. M. Thakare, "A study of human activity recognition using AdaBoost classifiers on WISDM dataset," The Institute of Integrative Omics and Applied Biotechnology Journal, 2016 Jan 1;7(2):68-76

[14]. K. Shaleena and S. Paul, "Data mining techniques for predicting student performance," in Engineering and Technology (ICETECH), 2015 IEEE International Conference on. IEEE, 2015, pp. 1–3.

[15]. M. Kumar, A. Singh, and D. Handa, "Literature survey on educational dropout prediction," IJ Education and Management Engineering, vol. 2, pp. 8–19, 2017.

[16]. M. Samner, E. Frank, and M. Hall, "Speeding up Logistic Model Tree Induction," In European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg.

[17]. J. R. Quinlal, "Bagging, Boosting, and C4.5", In AAAI/IAAI, Vol. 1, pp. 725-730. 1996.

Cite this article as :

Sumita Guddin, Dr. R. N. Yadawad, Dr. U.P. Kulkarni, "Stream Processing for Performance Analysis of Identifying Dropout Students utilizing Different Decision Tree Based Methods", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 4, pp. 143-149, July-August 2022. Available at doi : https://doi.org/10.32628/CSEIT228417
Journal URL : https://ijsrcseit.com/CSEIT228417