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

Samyukta Guddin[1], Puneeth G M[2], Shagufta Tangi [3] Abhilasha Naik[4] Sumita Guddin[5]

[1,2,3,4]UG Student, Department of CSE, GECK, Karwar, Karnataka, India
[5] Professor Department of CSE, GECK, Karwar, Karnataka, India

---***---

Abstract - Education paves the way for a person to live a secure, prosperous life. In a similar manner, the percentage of a nation's population with a higher degree of education may have an impact on that nation's progress. However, this number is decreasing as a result of early school abandonment. Furthermore, if a student cannot continue due to a dropout, the resources of the country are decreased. Despite the fact that the percentage of dropouts is steadily decreasing, it is still quite difficult for educational institutions to detect these students. The improvement of student performance is a school's top goal, thus it makes sure that all students graduate on time. Student dropout, however, is a substantial obstacle that adversely affects this aim. Finding out what causes dropouts is essential to determining a response. The root factors vary from student to student; some are related to workload and mental toughness. In this work, many decision tree (DT) approaches have been proposed and investigated.

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

1.INTRODUCTION

Education helps us study, pick up knowledge, and develop our skills. It is essential for both global and national growth. Education has a profound impact on our minds and personalities. It encourages us to have an optimistic outlook. Recently, modelling, understanding, and forecasting student performance and academic progress have attracted a lot of interest [1]. The educational system of any country greatly influences its progress. This is why a lot of academic institutions stress on-time graduation. In relation to

Major challenges include the lack of graduates and ways to improve student performance in educational institutions. The chance that a graduate will complete their degree on time may be affected by a number of factors, including financial limitations, a causal attitude, unanticipated life events, and others. Finding the youngsters by watching their behaviour takes some time. Numerous labour hours are required. In this case, dropout students can be located by using data mining techniques. From massive volumes of data, we can extract hidden information using a technique called data mining [2]. Data mining includes both descriptive and predictive elements. Supervised learning techniques are often predictive. Unsupervised learning, however, is more demonstrative [3]. For training purposes, the supervised learning technique provides input examples together with their class labels. We can make predictions about the test data by examining the data that has been provided. The approaches for detecting dropout students developed, implemented, and assessed in this study are based on Random Forest (RF), Logistic Model Tree (LMT), Decision Tree (CART), and AdaBoost Decision Tree (ABT).
Apache Kafka, an open-source stream platform, has been used to handle the data streams in real time.

The rest of the paper is organised as follows: II's emphasis will be on literary reviews. Section III discusses how to determine an attribute's importance. In Part IV, a variety of tree-based techniques are discussed. The synthetic dataset is succinctly described in Section V. Evaluation metrics were covered in Section VI. The tree investigated in Section VII serves as an illustration of the efficacy of tree-based approaches. The outcome of our effort is discussed in Section VIII.

2. LITERATURE SURVEY

Student dropouts lead the country to regress, and the educational system's resources are exhausted. Finding out which students have left is a challenging task. To target dropout youngsters for this, researchers have tried a variety of classification techniques. All IT-using firms employ digital data as the de facto currency in today's society. The capacity to analyse massive data quantities, or "big data," has arisen as a competitive advantage as a result of the global expansion in data. New waves of productivity, growth, and innovation are fueled by big data. The research [4–9] offers insights into a range of factors, such as student performance and different learning styles, that may have an impact on memory in STEM-related courses. [4] According to the research paper "Student Attraction, Persistence, and Retention in

STEM Programmes: Successes and Continuing Challenges," regarding IT studies in particular, there are many unfavourable presumptions and assumptions about the nature of the IT profession. These myths frequently discourage young people from pursuing IT jobs and programmes, and some even give up on IT (or majoring in IT) [5]. The authors of a recent study [6] want to evaluate how well several categorization techniques (naive Bayes, neural networks, support vector machines, decision trees, and random forests) perform for analytical students. Additionally, categorization rules have been devised. [7] provides a thorough examination of student accomplishment across a number of years. In order to identify pupils that don't reach the necessary levels, a number of classification techniques have been utilised, such as decision trees, decision tables, logistic regression, and naive bayes. The authors of a work [8] that examines several tree-based methodologies demonstrate that each tree performs better in each circumstance. They've come to the conclusion that choosing the right parameter is essential for knowledge discovery. The REP Tree and Decision Tree algorithms were compared by the authors of the study [9]. They provided examples of how the trimming algorithm facilitates accurate data analysis. They also concentrated on the effects of changes in accuracy and complexity on the REP tree. The authors of [10] (CART) have compared the efficacy of several techniques, including the Random Forest algorithm, Logistic Model Tree (LMT), and Classification and Regression Tree. In terms of prediction, Random Forest outperforms CART and LMT.

3. 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 wereemployed in our work are essential for figuring out how much the dependent property is worth. Equation1 can be used to compute the information gain.

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

Hence "c_i" refers to class. P(t) is the likelihood thatthe 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 $P(c_i/(1\text{-}t))$ 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, thedimensionality 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.

4.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 themfor 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 used using 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 forboth 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)

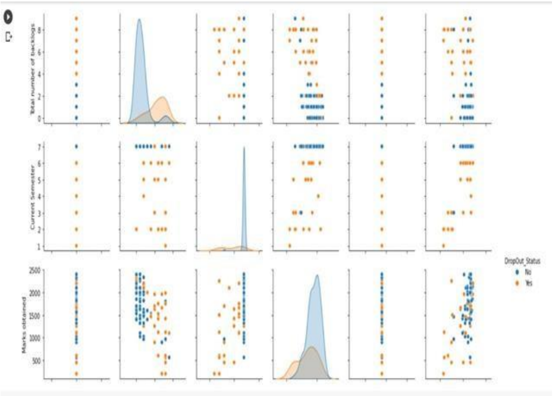
AdaBoost works best to improve a weak learner's performance. AdaBoost is a straightforward and organic extension of AdaBoost that Freund and Schapire presented for k>2 classes. M1. Every time a boosting approach is used, a weight is kept. When the weight is bigger, the classifier's impact over the learner is greater. By merging the entire set at the end, it improves the performance of weak learners[17].

5. DATABASE DESCRIPTION

It has been generated a dataset called "Student Performance Analysis." 9 attributes are included inthis dataset, including Name, Address, Semester, Current Semester, and Total backlogs, overall gradetotals, grades earned, number of working days, overallclass attendance, drop-out status, and other statistics. Instances in the dataset total 106. The dataset is usedto 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.



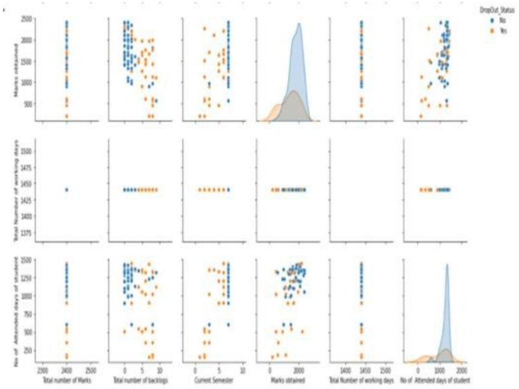


Figure 1 : Student performance analysis dataset Inform of pair plot

6. 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

$$Precision = TP / TP + FP \ \ldots (2)$$

$$Recall = TP / TP + FN \ \ldots (3)$$

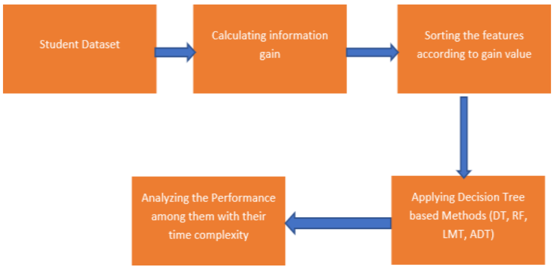
$$F1measure = 2 * Precision * Recall / \ Precision + Recall \ \ldots (4)$$

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


Student Dataset

Calculating information gain

Sorting the features according to gain value

Applying Decision Tree based Methods (DT, RF, LMT, ADT)

Analyzing the Performance among them with their time complexity 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.

$$MCC = TP * TN - FP * FN / \sqrt{}\,(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN) \ldots (5)$$

E. Root Mean Squared Error (RMSE)

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

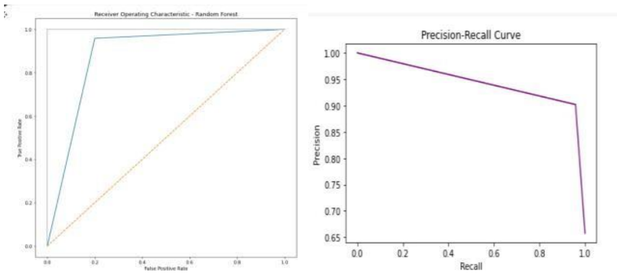
$$RMSE = \sqrt{}\sum_{t=1}^{n} (fi - y_i)^2 / n \ldots. (6)$$

7. 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

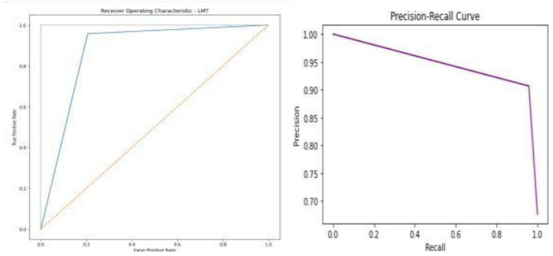


Figure 4: i) Receiver Operating Characteristic Logistic Model Tree, ii) precision _ recall _curve

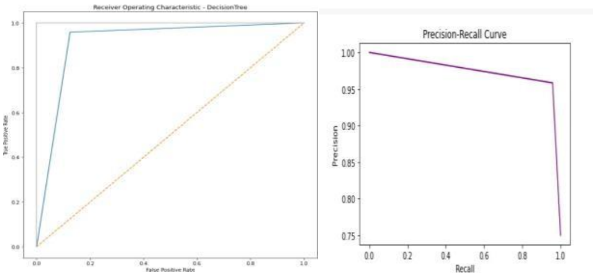


Figure 5: i) Receiver Operating Characteristic Decision Tree, ii) precision _ recall _curve

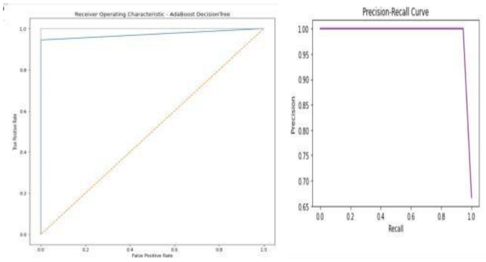
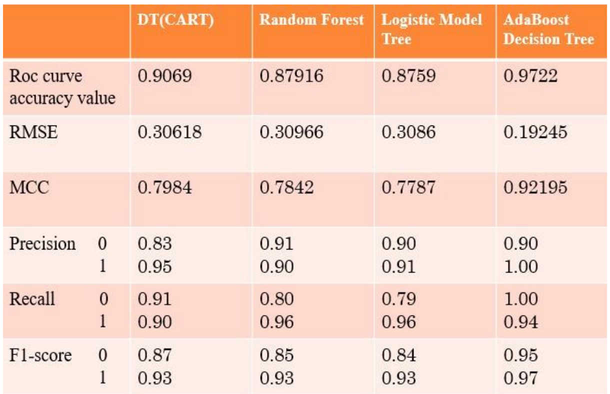


Figure 6: i) Receiver Operating Characteristic AdaBoost Decision Tree, ii) precision _ recall _curve

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.

8. CONCLUSION

This study compares various tree-based methods for locating dropout students. Utilising Kafka to stream data is a model for summarising the student dataset. To identify dropout pupils, important characteristics are calculated. The most important components are thought to have the largest gain. The efficacy of the techniques is evaluated using a variety of parameters. Examples of decision trees are the AdaBoost Decision Tree, Random Forest, and LMT. In the future, XG Boost, a contemporary tree-based method, can be compared to

9.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.