Machine Learning Algorithms to Predict Potential Dropout in High-Schools

Vaibhav Singh Makhloga, Kartikay Raheja, Rishabh Jain and Orijit Bhattacharya

July 21, 2020

MACHINE LEARNING ALGORITHMS TO PREDICT POTENTIAL DROPOUT IN HIGH-SCHOOL

Vaibhav Singh Makhloga[1], Kartikay Raheja[2], Rishabh Jain[3] * and Orijit Bhattacharya[4]

[1]Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi 110053, India

makhlogavaibhav@gmail.com

[2]Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi 110053, India

rahejakartikay99@gmail.com

[3]Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi 110053, India

rishab1300@gmail.com

[4]Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi 110053, India

orijit98@gmail.com

Abstract

In a developing country like India, the growth of its citizens and consequently the advancement of the nation depends on the education provided to them. However, the process of delivering education has been hindered by considerable dropout rates which has multiple social and economic consequences. Hence, it is crucial to find out ways to overcome this problem. The advent of machine learning and the availability of an immense amount of data has enabled the development of data science and consequently, its application in Education Institutions. Educational data mining enables the educator/teacher to monitor student requirement and provide the necessary response and counselling.. In this paper we use advance machine learning algorithms like logistic regression, decision trees and K-Nearest Neighbours to predict whether a student will drop out or continue his/her education. The accuracy of such models is calculated and studied. On the basis of the results it was found that ML techniques prove to be useful in this domain with random forest being the most accurate classifier for predicting dropout rate. Educational institutions can analyze which students may need more attention using this research as it's base, thus modifying teaching methods to achieve the end goal of 0% dropout rate.

Keywords - Educational data mining, Classification, KNN, Decision tree, Logistic Regression Predictive analysis, Dropout detection

1. Introduction

The growth in educational sector in India has accelerated over the past few years. There are around 12,50,775 govt schools in India and an estimated 339,000 private schools. The education sector is faced with a major concern and challenge in the form of high and increasing dropout rate in the senior secondary Level [17]. Transition from middle school to high school ie. classes 10th and 11th grade, is considered to be the most critical period when students show warning signs. Courses become intellectually demanding, peer groups are larger and students experience personal freedom[13]. Other factors such as Social & economic background, demographic and family background also influence a child's schooling [16]. Negative consequences of these student dropout are significant both to individual and society. There will also be an impact on the educational institution's reputation which may lead to lower ranking and reduced government funding. A lack of formal education in case of drop out students could restrict their economic well being later in life. The nation as well as society could suffer as drop outs are frequent recipients of welfare schemes and unemployed subsidies [14]. Policymakers, Educators and researchers have long considered student dropouts as a serious education problem that needs to be tackled.

The Base of this research paper is built upon the emerging research field of EDM. The analysis of educational data using statistics and exploration has proven to be insightful in understanding student behaviour [18]. The rapid development of machine learning models and their accuracy at Predicting future trends which makes it a promising tool to tackle a wide range of real life problems [19]. The scope of this research is to combine the above two fields and determine it's accuracy in predicting potential drop-outs early so that schools can intervene at the right time.

In this paper, we apply Machine learning and build :

- Predictive models to calculate the precision and accuracy of classification algorithms such as Logistic Regression (along with gradient descent) , KNN, Decision Trees in predicting school drop-outs. The educational data is formed by combining datasets collected from National Informatics centre (NIC) and Ministry of Electronics

- EDM methodologies are used to find correlation between various factors that lead to student dropout based on graphs generated by algorithms.

- The result of the classifiers are analysed using the precision metrics and to prevent unbalanced classification ROC curve is used

The selection of the dataset and it's features plays a crucial role in our research. They will take into account various factors like

- Behaviour (eg. number of suspensions)

- Performance (eg Examination and test marks),

- Facilities (eg. Internet access, no of toilets, teaching staff) among others.

- Family background (eg. caste and Guardians)

The result has shown that Decision Tree algorithm has been able to predict drop-out with high Accuracy. This shows that the usage of machine learning to the increasing drop-out problem could be useful. A Warning system can be developed based on this research built upon machine learning algorithm which can inform the school authorities about potential drop out behaviour so that there can be an early intervention and special attention can be given to those at risk.

The following section describes in brief the Background and related work. A brief description of Machine Learning techniques used in our study is section 3. Section 4 presents details about the dataset. The experiment along with training and testing phase is mentioned in Section 5. Section 6 discusses the result and has further analysis. Finally, Section 7 concludes our research.

2. Background and related work

Kotsiantis, Pierrakeas and Pintelas [1] used machine learning techniques to deal with the problem of student dropout in universities providing distance education. Attempts have been made by the researchers to develop an appropriate learning algorithm for prediction of student's dropout.

Yukselturk, Ozekes, Türel [2] examined the prediction of dropouts using data mining approaches in an online education program. Data was collected through online questionnaires which included variables such as gender, age, educational level, self-efficacy, previous online experience, online learning readiness etc. To classify data, four data mining approaches were used based on k-Nearest Neighbour, Decision Tree, Naive Bayes and Neural Network. Moreover, the Genetic Algorithm was used to find the significant determinant factor(s) amongst the factors mentioned above.

Aulck, Velagapudi, Blumenstock and West [3] described initial efforts to model student dropout using the most extensive dataset on higher education attrition that tracks over 32,500 students' demographics and transcripts. Their model highlights several early indicators and accurate dropout prediction, even based on a single term academic transcript data.

Suh, Suh & Houston [5] attempted to identify the critical factors of at-risk students: those with low-grade point averages, those who had been suspended, and those from low socioeconomic backgrounds. The author carried out a logistic regression analysis of the data, obtained from the National Longitudinal Survey of Youth−1997, which indicated that student dropout rates were affected differently by students' membership in the three at−risk categories.

3. Machine learning techniques

3.1 Logistic regression

As per Klein [6] and Menard [7], it is a model based on statistics that uses logistic functions in its basic form to model a binary dependent variable, even though many more complex extensions are existing. It estimates the parameters of the logistic model only, which means a form of binary regression. We have used logistic regression to classify the pattern of students into two different classes where one is dropping out from any particular course, and the another continues the course initiated.

3.2 KNN Algorithm

K-nearest neighbours algorithm is easy to implement a supervised machine learning algorithm. KNN is used to solve regression and classification problems. It has no model rather than storing the entire dataset. It relies on labelled input data to learn a function that produces an appropriate output when given new unlabeled data. Soucy, Mineau [8] and Yigit [9] proposed KNN algorithms for text categorization. When KNN is used for classification, the output can be calculated as the class with the highest frequency from the K-most similar instances.

3.3 Random forest

Random Forest model includes a simple decision tree which is a weak learner and give us the output at average, or the output is the majority vote of the decision tree. Random Forest model is known for various advantages such as robustness against noise, quick learning and easy setting of hyperparameters. Liaw and Weiner [11] used Random Forest for implementing classification and regression. We have used Random Forest to classify our given set of data into a more straightforward form. Logistic Regression and Random Forest both work as classifier but at different accuracies together gives us a trained data which is sent to the Scikit-learn for the construction of prediction model.

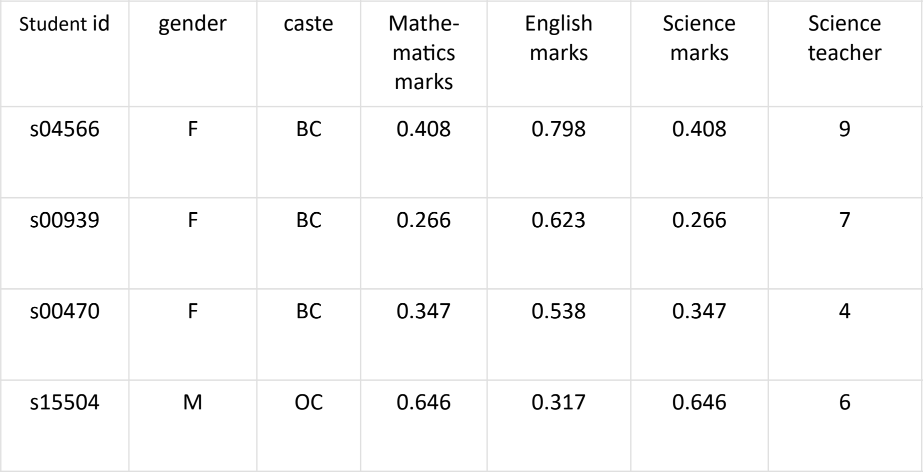
3.4 ROC Curve

ROC curve is a graphical display of sensitivity (TPR) on the y-axis and (1 – specificity) (FPR) on the x-axis for varying cut-off points of test values. This is generally depicted in a square box for convenience and its both axes are from 0 to 1. The area under the curve (AUC) is an effective and combined measure of sensitivity and specificity for assessing inherent validity of a diagnostic test. Maximum AUC = 1 and it means diagnostic test is perfect in differentiating diseased with non-diseased subjects. The investigation done by Bradley [12] supported the use of ROC in our research.

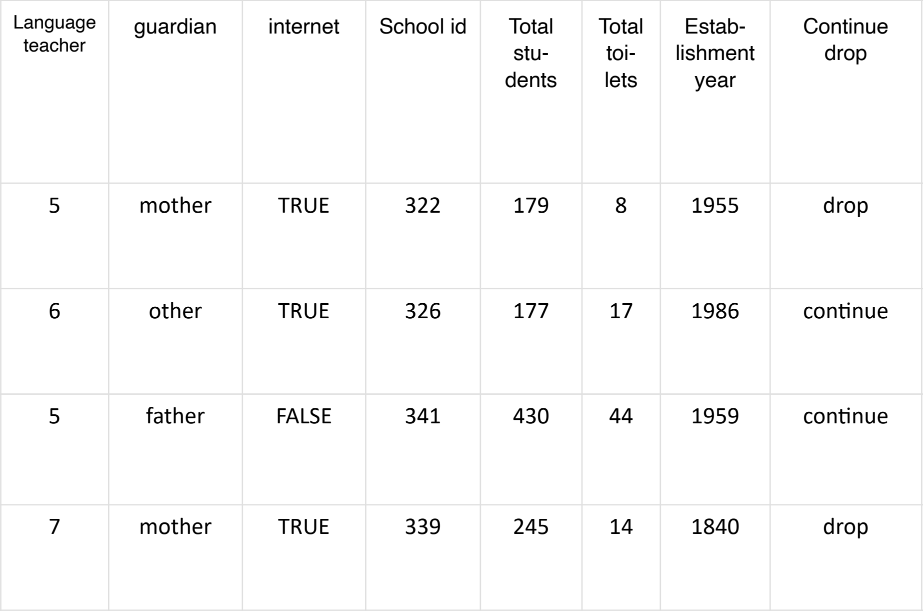
4. Dataset

The data was collected by the National Informatics Centre (NIC), Ministry of Electronics & Information Technology, Government of India.The National Informatics Centre (NIC) is an attached office under the Ministry of Electronics and Information Technology (MeitY) in the Indian Government, established in 1976 under the Planning Commission of the Indian government.

It is a representative of the quality of High school education provided in our country in relation to student dropouts in 2011-12. The data contains some features about the conditions in the school and the performance of the student to determine whether he/ she drops out or not. It features conditions provided at the school in the following categories total toilets, availability of science and language teachers, internet availability and establishment year of the school. Student performance is judged using marks in science, mathematics and language also features like caste, gender and present guardian of the student are also taken into account to determine whether he/ she drops out. In this paper, we considered factors other than the student's performance, such as socio-economic conditions, age, infrastructure barriers of the students while extracting the data from NIC portals.



Student id	gender	caste	Mathematics marks	English marks	Science marks	Science teacher
s04566	F	BC	0.408	0.798	0.408	9
s00939	F	BC	0.266	0.623	0.266	7
s00470	F	BC	0.347	0.538	0.347	4
s15504	M	OC	0.646	0.317	0.646	6



Language teacher	guardian	internet	School id	Total students	Total toilets	Establishment year	Continue drop
5	mother	TRUE	322	179	8	1955	drop
6	other	TRUE	326	177	17	1986	continue
5	father	FALSE	341	430	44	1959	continue
7	mother	TRUE	339	245	14	1840	drop

4.1 Data processing

Due to high skewness in data towards the number of students continuing the education, bias was introduced in the data for the prediction model. Initially, the data contained information about 19000 students out of which we sampled 1800 values as a representative of the data to create asymmetry between dropouts and students continuing the education. After some data cleaning process, 1763 data values were finalized due to presence of some null values.

4.2 Numerical Data:

Other than cleaning the data and dropping the null values, some feature scaling was also done to bring certain features values on the same scale.

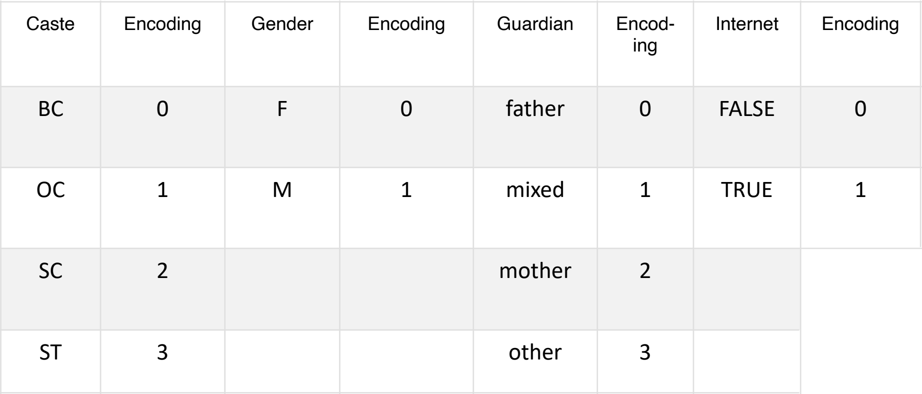
$$z = x\text{-}u/\ r \qquad (1)$$

Where μ= Mean Value σ or r= Standard Deviation

We scaled two features total_students and total_toilets to get them to the same scale so that the value of one feature should not dominate over the others and hinder the performance of the learning algorithm.

4.3 Categorical Data:

Features like gender, caste, guardian and internet had categorical values, so they were encoded using a label encoder to integer values.



Caste	Encoding	Gender	Encoding	Guardian	Encoding	Internet	Encoding
BC	0	F	0	father	0	FALSE	0
OC	1	M	1	mixed	1	TRUE	1
SC	2			mother	2		
ST	3			other	3		

Encoding will enable these columns to be used as features in our classifier.

5. Experiment and Result

5.1 Training Phase:

Data were divided into training and testing parts with training data size of 1188 values with 11 features and one label. While testing data size was 586 values with 11 features and one label. This split was done by shuffling the data to get rid of any patterns in the data format that would bring bias or cost us in reduced accuracy. This training data was fed into various classification model for training the respective models, whereas the purpose of testing data set was to test the model for overfitting or any bias during the prediction phase.

Logistic Regression model gave a training accuracy of 66.07% and a testing accuracy of 64.86% using L2 regularisation for a penalty. Regularisation is a method of avoiding overfitting of classification models by penalizing high valued regression coefficients. L2 regularisation adds penalty equal to the square of the magnitude of coefficients

Confusion Matrices using heat maps were plotted for each model to obtain following values:

- True Positive (TP) : Observation is positive and prediction is also positive.
- False Negative (FN) : Observation is positive but prediction is negative.
- True Negative (TN) : Observation is negative and prediction is negative.
- False Positive (FP) : Observation is negative but prediction is positive.

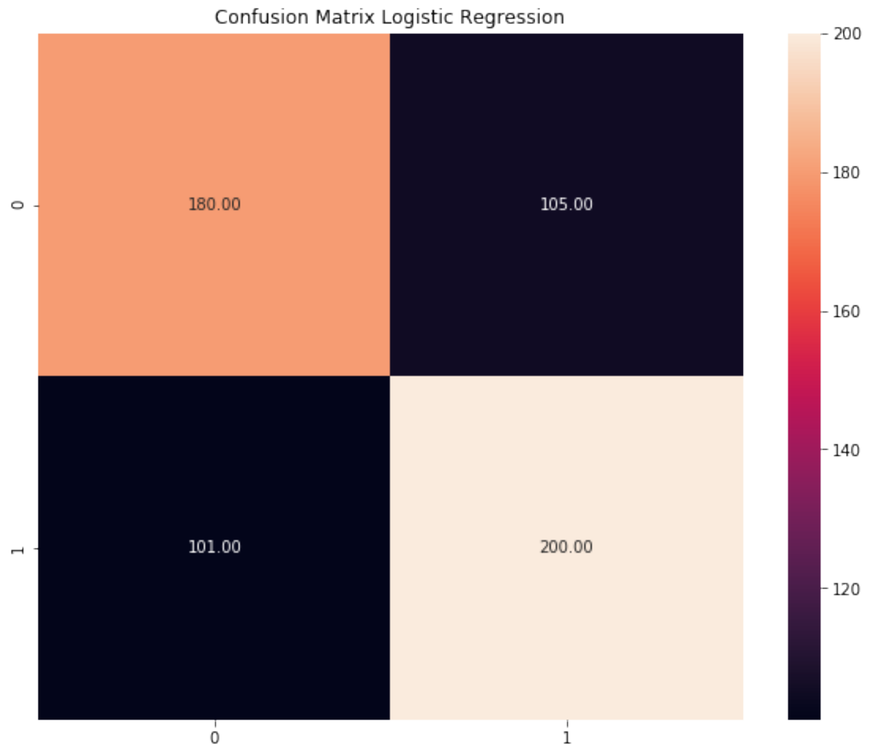


Fig1:- Using Logistic Regression we get, TP=180, FP=105, TN=200, FN=101

Regression Coefficients :- (1.23971592, 0.0410947 , -0.2475107 , 0.60797969, -0.10031828, 0.03504581, -0.01807157, 0.31297533, -0.8502106 ,2.8747373 ,-0.8502106)

In Decision Tree, the feature science_teacher has the maximum gain ratio and which has made it the starting node and the most effective attribute for the initial split. For a student to be classified as a dropout, the essential features according to the generated decision tree are science marks, English marks, number of language teachers and caste. Gini index or coefficient was used for the statistical analysis.

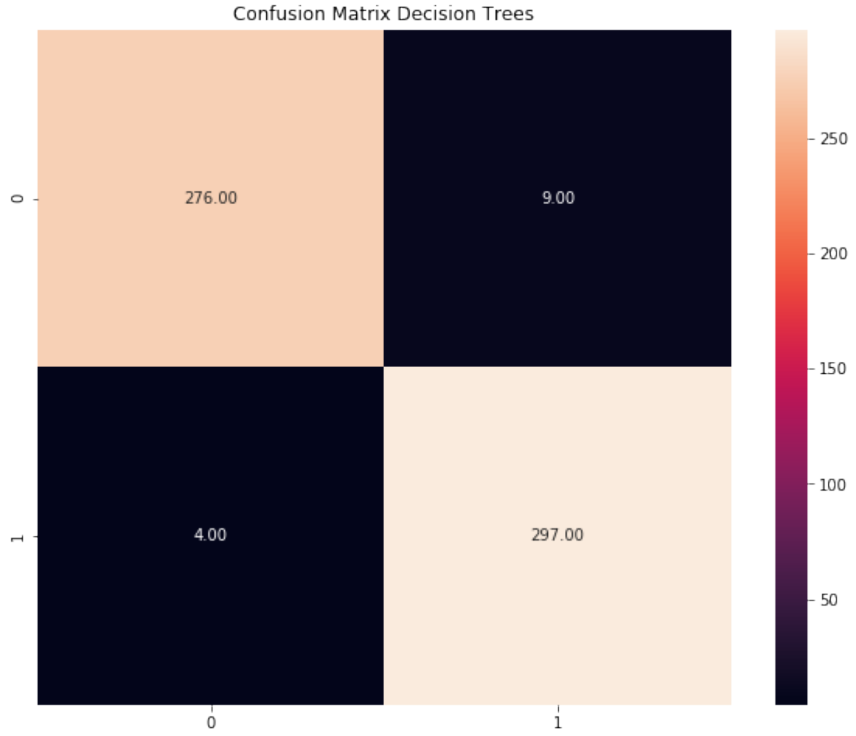


Fig 2 :- Using Decision Trees classifier we get,TP=276, FP=9, FN=4 and FP=297

KNN classifier gave us an accuracy of 92.2% the training set and an accuracy of 88.39% on the test set. The prediction was based on five nearest neighbours and a Minkowski metric. The Minkowski distance is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance.

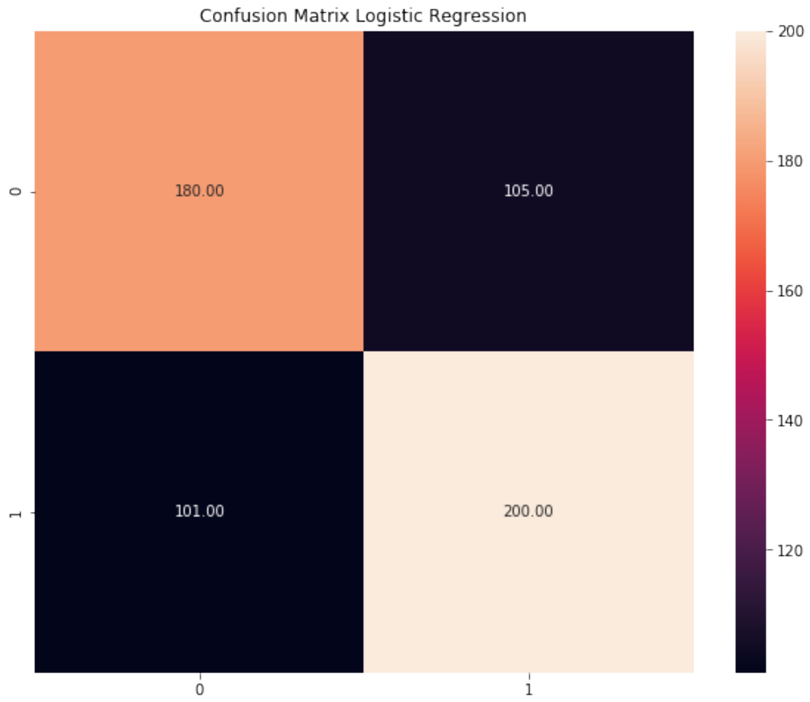


Fig 3 :- Using KNN Classifier we get, TP=220, FP=65, FN=3

& TN=298

5.2 Method of analysis

The entire data was divided into training and testing set in the ratio 4:1, and a Classification metric was generated to get the prediction result on the test data. Precision rate, recall rate, F-measure and the overall accuracy rate were used as indicators to get the effectiveness of the predictions.

Precision defines the percentage of samples with a specific predicted class label belonging to that class label.

$$\text{precision}=TP/(TP+FP) \tag{2}$$

Recall defines the percentage of samples of a particular class which were correctly predicted as belonging to that class. The f1 score is defined as the harmonic mean of precision and recall and is a far better indicator of model performance than precision and recall (usually).

$$\text{recall}=TP/(TP+FN) \tag{3}$$

6. Discussion

Three classification techniques were applied to the dataset to build the perfect model. These techniques are KNN, Logistic regression and Decision Forest. After the preprocessing and preparation of raw data into a usable format, we applied the three algorithms on it and got the accuracy score. Decision Tree was successfully able to predict the correct Y-Label i.e Drop-out approximately 98% of the time compared to basic algorithms such as Logistic Regression which was correctly able to predict only 66% of the time. KNN showed some promise with selecting 5 as nearest neighbour. Some fine tuning of the algorithm and it's parameters may lead to improved results. Being a complex classifier decision tree was able to perform better. It was seen that it took 16 levels of splits to get the desired accuracy. Levels were reduced so as to not overfit the data. We try to analyze the ROC curve visually. ROC curve is used here as a diagnostic tool to check if there are any imbalanced classification (i.e. a negative case with the majority of examples and a positive case with a minority of examples). It will also help us to compare the performance of the used classifiers using AUC (area under the curve).

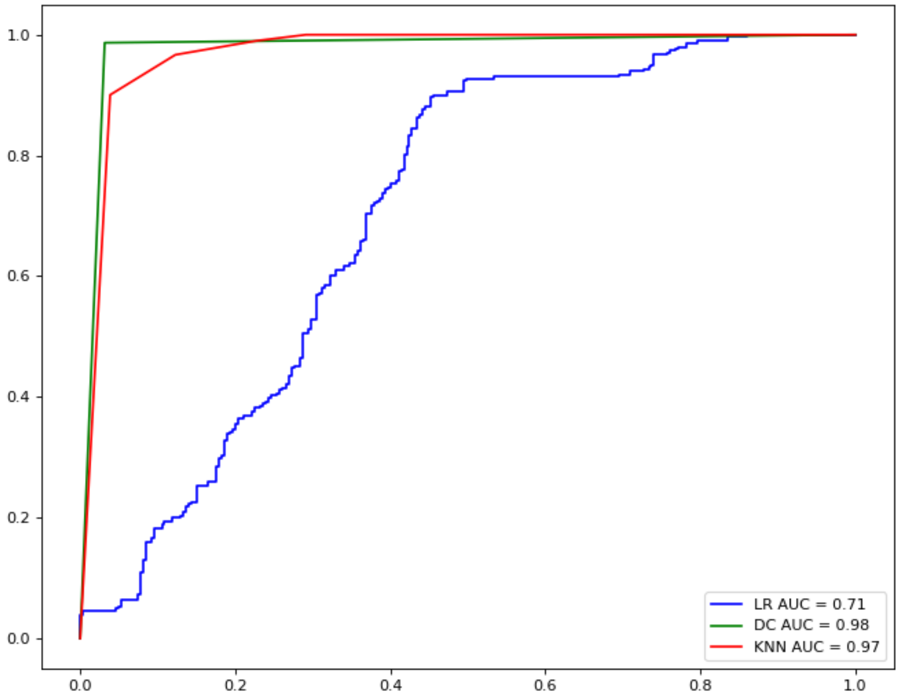
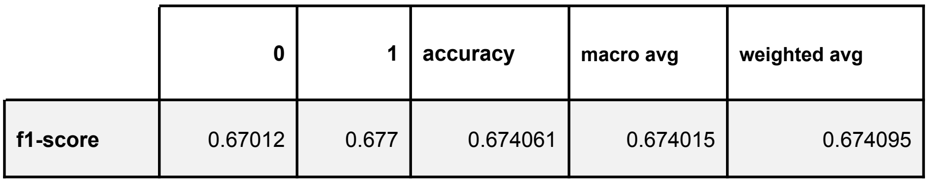


Figure 4 :- ROC curve

After studying the combined ROC curve, it is clear that the random forest gives the maximum accuracy in predicting dropouts while logistic regression has the least accuracy. AUC of KNN, Logistic regression and Decision tree is as follows: 0.97, 0.71, 0.98.

The following tables show the classification metrics for each Logistic Regression, Decision tree and KNN. They have precision values like 0.664, 0.987 and 0.847, respectively.

Logistic Regression Classification metrics:



	0	1	accuracy	macro avg	weighted avg
f1-score	0.67012	0.677	0.674061	0.674015	0.674095

precision	0.66438	0.683	0.674061	0.674029	0.674226
recall	0.67595	0.672	0.674061	0.674099	0.674061
support	287.000	299.0	0.674061	586.0000	586.000000

KNN Classification metrics:-

	0	**1**	**accuracy**	**macro avg**	**weighted avg**
f1-score	0.8745	0.903	0.890785	0.888917	0.889212
precision	1.0000	0.8236	0.890785	0.911846	0.910041
recall	0.7770	1.0000	0.890785	0.888502	0.890785
support	287.00	299.00	0.890785	586.00000	586.000000

Decision Tree classification metrics:-

	0	**1**	**accuracy**	**macro avg**	**weighted avg**
f1-score	0.9877	0.9883	0.988055	0.988047	0.988053
precision	0.9929	0.9834	0.988055	0.988201	0.988103
recall	0.9825	0.9933	0.988055	0.987945	0.988055
support	287.00	299.00	0.988055	586.00000	586.000000

7. Conclusion

This study lays the preliminary foundation for building a system for detecting students that are at risk of dropping out. We examined the machine learning techniques such as Decision trees, Logistic Regression and KNN to students database and investigated their performance This research shows that machine learning algorithms are efficient in predicting dropouts and similar methods can be adopted by every school for early identification of at-risk students

The approach implemented here can detect signs of student's disengagement from the learning environment for different age groups on different education levels. Proof of concept implementation shows a certain level of viability of this approach.

Moreover, the ranking of contrastive variables obtained by this approach would help to determine which variable affects the learning process the most at a given point of time. By monitoring the critical variables constantly, the learning process is made more adaptive to the student and planning of the institution's curriculum can be considered based on this information in order to minimize the dropout rates.

References

1. Kotsiantis S.B., Pierrakeas C.J., Pintelas P.E. (2003) Preventing Student Dropout in Distance Learning Using Machine Learning Techniques. In: Palade V., Howlett R.J., Jain L. (eds) Knowledge-Based Intelligent Information and Engineering Systems. KES 2003. Lecture Notes in Computer Science, vol 2774. Springer, Berlin, Heidelberg

2. Yukselturk E, Ozekes S, Türel Y (2014) Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. European Journal of Open, Distance and E-Learning 17:118-133. doi: 10.2478/eurodl-2014-0008

3. Aulck, Lovenoor S., Nishant Velagapudi, Joshua Evan Blumenstock and Jevin West. “Predicting Student Dropout in Higher Education.” *ArXiv* abs/1606.06364 (2016): n. pag.

4. Marius Kloft, Felix Stiehler, Zhilin Zheng, Niels Pinkwart, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 60–65, October 25-29, 2014, Doha, Qatar. c 2014 Association for Computational Linguistics.

5. Suhyun Suh, Jingyo Suh & Irene Houston, Volume 85, Issue 2, Pages: 131-255, Spring 2007, DOI: https://doi.org/10.1002/j.1556-6678.2007.tb00463.x

6. David G. Kleinbaum, Mitchel Klein, Logistic Regression: A Self-Learning Text, Springer, New York, NY, 2010, DOI: https://doi.org/10.1007/978-1-4419-1742-3

7. Scott M., Applied Logistic Regression Analysis, second ed., SAGE, 2001, pages:1-33 https://books.google.co.in/books?id=JbVIDwAAQBAJ

8. P. Soucy and G. W. Mineau, "A simple KNN algorithm for text categorization," *Proceedings 2001 IEEE International Conference on Data Mining*, San Jose, CA, USA, 2001, pp. 647-648.

9. H. Yigit, "A-weighting approach for KNN classifier," 2013 International Conference on Electronics, Computer and Computation (ICECCO), Ankara, 2013, pp. 228-231.

10. K. J. de O. Santos, A. G. Menezes, A. B. de Carvalho and C. A. E. Montesco, "Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout," *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, Maceió, Brazil, 2019, pp. 207-208.

11. Liaw, A. and Wiener, M. (2002) Classification and Regression by Random Forest. R News, 2, pp. 18-22

12. Bradley, Andrew P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), pp. 1145-1159.

13. M., Sateesh & Sekher, T V. (2014). Factors Leading to School Dropouts in India: An Analysis of National Family Health Survey-3 Data. International Journal of Research & Method in Education. 4. 75-83. 10.9790/7388-04637583.

14. Edward J. McCaul, Gordon A. Donaldson Jr., Theodore Coladarci and William E. Davis
The Journal of Educational Research
Vol. 85, No. 4 (Mar. - Apr., 1992), pp. 198-207

15. Rumberger, R. W. (2001). Why Students Drop Out of School and What Can Be Done. UCLA: The Civil Rights Project / Proyecto Derechos Civiles. Retrieved from https://escholarship.org/uc/item/58p2c3wp

16. Kominski, R. Estimating the National High School Dropout Rate. *Demography* **27,** 303–311 (1990). https://doi.org/10.2307/2061455

17. C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601-618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

18. Pat Langley and Herbert A. Simon. 1995. Applications of machine learning and rule induction. Commun. ACM 38, 11 (Nov. 1995), 54–64.