International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Student Dropout Prediction Using Machine Learning Techniques

Haarika Dasi[1], Srinivas Kanakala[2]

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract: In recent years, the number of students that drop out of school has substantially increased. Many educational institutions or universities have been threatened by the high percentage of students who abandon a registered course, the common study subject in the learning analytics field is the early and accurate prediction of Student Dropout depending on the available educational data. Despite the volume of completed study, there has been little advancement and this trend has persisted across all levels of educational data. Although many features have already been studied, it is still unclear which features may be used with various machine learning classifiers for the forecasting of student drop out. A major objective of this research is to highlight the importance of understanding and gathering data, emphasize the limitations of the available educational datasets, compare machine learning classifier performance, and demonstrate that, if performance metrics are carefully taken into consideration, even a limited set of features teachers can use to predict student dropout in an e-learning course can be accurate. Four academic years' worth of data was evaluated. The features chosen for this investigation worked well in identifying course completers and dropouts. On unobserved data from the upcoming school year, the prediction accuracy ranged between 92 and 93%. Along with the commonly employed performance measurements, the homogeneity of machine learning classifiers was compared and studied in order to mitigate the effect of the small dataset size on the high performance metrics values. The outcomes demonstrated that a number of machine learning methods might be used successfully to analyse an academic data which is of tiny size. this can result in a page being rejected by search engines. Ensure that your abstract reads well and is grammatically correct.

Keywords: Support Vector Machine (SVM), Neural Networks (NN), Naïve Bayes (NB), Dropout prediction, Decision Tree (DT), Machine Learning (ML), Random Forest (RF), Logistic Regression (LR)

1. Introduction

The idiom "dropout" refers to a student who drops out of school, college, or university without completing the required course. Every year, approximately a million students drop out of educational institution without finishing their courses. Individual and institutional elements both influence exit rates. Data technology is always evolving, and this has a big impact on both education & business. The Govt., of India, both the state and the federal, places a high value on education each year through promoting compulsory education, establishing educational funds for both teaching and research, and more. However, a lot of students continue to drop out of school, which prevents them from graduating or delays their graduation. This has a severe impact on the future commercial and industrial workers as well as the students' institutions, which lose money and time as a result.

If we can recognize the early signs of dropout and pay attention to them, the rate of student abandonment will be reduced. Therefore, several research are being conducted in numerous nations to examine the variables influencing student performance in higher education. There are nine categories of characteristics that affected students' performance in the earlier studies: Students' treasured grades and academic achievement, their e-learning activities, their social demographics, their social information, their instructor's qualities, their course qualities, their evaluations of the course, their environment, and their experiences are just a few examples. Researchers have different definitions of dropouts, but regardless, if any institutions lose any student in any way, the retention rate is low. Any school retention strategy must be successful if it is to prevent dropouts from occurring. Identification of the factors affecting dropout rates is essential in order to respond to them, develop interventions to help students stay in school, and reduce the risk of the dropout rate rising. Institutions, the general public & students are all harmed by student dropout in schools. Institution attrition wastes both private and public resources, even when the student improves academically before leaving. Also, dropouts can cause feelings of inadequacy and social stigma in addition to financial loss. On the other hand, failure can adversely affect pupils as well as schools. When a student can't complete his or her education within the allotted study period, dropout occurs. A dropout student's knowledge and skills in his or her discipline are significantly less than those of a retention student, causing the institution's quality to suffer. It is important to note that despite the significance of the topic, a large amount of misinformation exists regarding the causes, symptoms, and effects of dropout, as well as how to minimize student attrition. In a learning environment, a predictive model is used to pinpoint the main reasons for dropouts using classifiers of data mining.

[1] Department of Computer Science and Engineering, VNR VJIET– 500090, INDIA
ORCID ID: 0000-3343-7165-777X
[2] Department of Computer Science and Engineering, VNR VJIET– 500090, INDIA

The Machine Learning (ML) has gained attention in recent years as a potential solution to the problem of student dropout. The reason for this is ml algorithms help in identifying children who are at risk and arrange interventions in a timely manner. This project will allow the school/institutions to decide how to proceed and create better plans that will prevent academic risk as a phenomenon and its effects, thus reducing the chances of withdrawal. This suggested methodology in this situation uses machine learning algorithms to forecast the student dropout rate based on a few variables.

2. RELATED LITERATURE

The literature has used a variety of ML algorithms, according to the results of the survey. Most of these algorithms, on the other hand, were created and tested in industrialized countries. As a result, ML has been applied in few studies to solve this problem in developing countries. Dropout rates have been investigated from several perspectives. Several ML techniques have been used and implemented in several research articles, and a few of them have been researched and described in the points below as a literature survey.

In this paper [1] by Mercy Paul Selvan et.al, in the realm of education, machine learning approaches are very important. Here, they've projected a victimization model that may be able to forecast dropout and provide a response to the prediction using decision tree (DT) algorithm. The accuracy observed in this model is 97.67%. Many factors were considered for the cause of dropout, such as family history, money, and social activities, which added to the predictability of the status.

The main goal of this paper [2] by Janka Kabathova et.al. is to emphasize the importance of data collection, highlight the limitations of available educational data datasets, and compare the performance of various ML classifiers such as these include SVM, neural networks (NN), LR, RF, and NB. With a 93 percent success rate, naive bayes achieved the maximum accuracy possible for the evaluation of the dataset, which includes variables access, tests, assignments, exams, and projects collected during four academic years.

In this paper [3], B. johannes et al. present an Early detection of students (EDS) at Risk model, using data of administration from German universities and ML methods to predict dropouts such as NN, bagging RF, and meta-algorithm AdaBoost were utilized in the prediction of overall student dropout, with AdaBoost achieving a high accuracy of 94.76 percent.

2007 and 2012, Kemper [4] anticipated dropout rates of students at Karlsruhe Institute of Technology (KIT Industrial)'s Engineering. 2,557 students successfully finished their degrees, while 623 dropped out. Logistic regression and decision trees were employed in this work to classify the data. When only the dataset of 1st semesters was trained, both techniques produced predictions with an accuracy of more than 83%; when the three semesters' worth of data was trained, the accuracy increased to up to 95%.

This study [5] helps to answer current global concerns about student dropouts. It allows colleges to identify students on the verge of dropping out and improve retention rates and educational quality by employing four machine learning techniques: RF, NB, DT, and LR. Algorithms like these are used to calculate Divide strength and capabilities into two categories: fail or pass. When compared to other algorithms, the prediction results demonstrate that LR has a high accuracy of 98 percent.

By utilizing data categorization approaches like as decision trees, Bayesian belief networks, and LR with 2,272 undergraduate students, Rawengwan et.al [6] evaluated variables affecting students' academic performance in the private colleges. The findings indicated that there are nine important variables: income, marital status, loan, type of school, educational background, sex, mother's occupation, student address, & school level. The accuracy of the Bayesian belief network, DT, and LR models is 78.9%, 85.3% and 78.5%, respectively.

The purpose of this work [7] is to build the best accurate forecasting model possible utilizing just academic data and ML methods such as GB, RF, and SVM. Where RF and GB have demonstrated 96% accuracy in forecasting student dropout using a dataset gathered over two academic years and based on numerous characteristics such as degree name, dropout, success rate, marks median, mark (entrance exam), and so on.

Anjana Pradeep [8] used a variety of classification algorithms, including Induction Rules and Decision Trees, to predict student dropout & failure in Kerala's M.G. university, India, from 2013-2018. The results of the AD Tree decision with the better qualities showed accuracy of 92%.

This study's [9] main goal is to make use of statistics to forecast and identify students who are likely to fail semester exams. Students' transcript data was evaluated, which includes their CGPA and grades in all university-level courses. The NB, NN, SVM, and DT classifier were among the ML methods employed in this analysis. A comparison of the accuracy results says that NB fared well in prediction with 60.01 percent accuracy.

This paper [10] aim was to create a system to identify new elements that could predict dropout, with the students, institutions, academic setting, and social and economic environment as the dimensions of analysis. Additionally, using LR, SVM and DT to see if the hypothesized factors are related and/or can help predict dropout at Ecuadorian colleges. When compared to LR and SVM, DT performed well, with an accuracy of 98 percent.

In this research [11], a data mining classification technique is used to propose an AI based system for predicting the purpose of college dropouts. This work employs machine learning algorithms such as Enhanced ML Algorithms (EMLA), NDTREE, and Decision Stump. The proposed methods are tested on a benchmarked dataset of students who have dropped out. According to the classification data, EMLA has a 78.31 percent accuracy, NDTREE has a 70.02 percent accuracy, and Decision Stump has a 30.71 percent accuracy. Other data mining techniques will be evaluated by the authors in future research.

The models in this study [12] by Marcell Nagy et al. are based on data (personal details, secondary school performance) from Budapest University of Technology and Economics who entered between 2010 and 2017 and completed their UG studies by graduation or dropping out. Following FE and FS, a variety of classifiers, including GB Trees, DT, NB, KNN, and DL were trained with various response situations. The best models, GB trees, and DL had AUCs of 0.808 and 0.811, respectively, were checkedapplying10-fold cross validation.

3. Methodology

Through the use of classification algorithms, this study seeks to examine the connections between students' personal data, academic histories, and academic standing from past semesters. Then, models are developed to forecast whether they would drop out of school or not in the following semester. The process diagram framework for our study is shown in Fig 3.1.

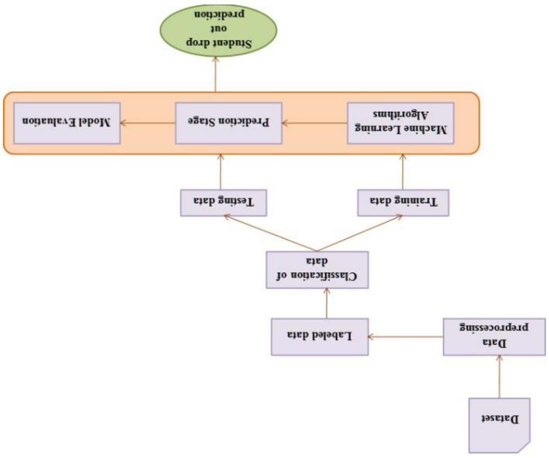


Fig 1: Process diagram Framework

A. Dataset

The online machine learning repository provides us with the dataset containing more than 261 student records both in the training and testing dataset with 10 parameters. The dataset contains information of the students in the form of exam, access, test grades, project, assignments, result points, tests, result grades, graduate project grade, and year. We are only concerned with few parameters which affect the drop out value.

B. Data Collection

In this phase, we gathered dataset from the Kaggle to evaluate and compare the performance of predictive machine learning techniques used in this project. In this experiment, data for independent features like exam, access, test grades, project, assignments, result points, tests, result grades, graduate project grade, and year were collected from several institutions in the form of a ".csv" file which results in the prediction of the dependent feature i.e., dropout or no dropout. We are concerned mainly with the three parameters like access, results and assignments to predict student drop out.

C. Pre-Processing of the Data

1.) **Data Cleaning & Transformation:** One of the crucial steps in the pre-processing of data stage is data cleaning. In this phase, the undesired data is eliminated, and the missing values or NA values are rectified. We then remove a few inaccurate and outlier data points that could lead to mistakes in our prediction models.

2.) **Feature Selection:** Access, assessments, and tests were chosen as significant elements that influence the dropout outcome in this experiment based on the significance of the factors for drop out prediction.

   a. Accesses are the total count of a student saw a course throughout the observational period.

   b. Assignments indicate the overall score from all the actions that were evaluated throughout the observation time.

   c. The results of the tests taken during the semester are the total of the results from the midterm and final exams.

As a final point, the variables access, assignments, and tests were selected because their importance lies in making predictions as early as possible so that intervention can be initiated. It must be emphasized that only the total number of partial tests conducted throughout the observation period was included. In addition to the two variables previously described with the number of accesses & good correlation (tests and projects), may be a more appropriate feature for prediction. The presence or frequency of access may allow us to identify students who do not complete the course in a timely manner, thus making them more likely to fail.

D. Model

Various machine learning methods are used to the dataset during the modelling process. A model is created based on students' current actions and successes in order to predict learner failure and performance in the future. An algorithm or classifier can resolve this typical classification problem by determining whether a student can complete the course.

To find a good predictive model, one can utilize one of the many various Machine Learning categorization techniques available today. The needs coming from the paper's primary goal and related works were taken into consideration, and the most widely used classifiers were ultimately used:

1) **Neural network:** This is a paralleled processing method that maximizes predicted accuracy by utilizing the structure and operations of the brain. Data is inputted into a deep learning NN, which subsequently analyses the data to produce the output across multiple layers.

2) **Logistic Regression:** In order to understand the data and characterize the relationship between an independent variable & a dependent variable that could be ordinal, interval in nature, nominal, or one method of prediction analysis called logistic regression is used. Finding the best model to explain the correlation between a group of independent factors & Boolean characteristic of interest and is the aim of this machine learning model, because the result's dependent variable is a binary variable.

3) **Naïve Bayes:** This classifier is a probabilistic straightforward classifier based on the theorem of bayes with stringent feature independence conditions.

4) **Decision Tree:** Decision trees are frequently utilized to solve classification & regression issues. The overarching objective of using decision tree is to construct a model which produces rules of decision from evaluating data sets and forecasts target classes. The decision tree adopts a structure of tree consisting of branches, roots, and leaves. In contrast to leaf nodes, which represent class labels, internal nodes represent decision-making qualities. Compared to other categorization strategies, the DT approach is easier to comprehend.

5) **Support Vector Machine:** It is a linear regression & classification model that may be applied to both nonlinear & linear problems. The algorithm divides the input into categories using a hyper plane. The value of each characteristic is represented by the value of a particular coordinate in this model, and in n dimensional space of points every item of data will be displayed as a point.

6) **Random Forest:** The Random Forest method generates a large number of ensembles of Classification & regression decision trees. A number of trees of decision are printed using a randomly chosen subset of the training datasets. The use of many decision trees increases the accuracy of the results. The algorithm allows missing data and has a relatively short runtime. Random forest is used to randomize the algorithm rather than the training data set. The sort of class that decision trees produce is the decision class.

The main goal of this paper is the comparison of the numerous prediction performances indicators of each used classifier in order to choose the most appropriate prediction model. The grid search method, which is frequently used to determine the best parameter settings, was utilized to modify hyper parameters.

4. Model Evaluation

It is challenging to construct an exact prediction point that is acceptable for picking out students who are at danger of dropping a course when the dataset is less. Data that was split into training and testing sets was used to compare the classification techniques stated above. To determine the confidence intervals for each of the performance indicators listed below, testing and training sets were selected at random:

1.) **Cross Validation:** This is a Statistical model of validation approach called cross-validation evaluates the generalizability of a study' findings to a different dataset. The accuracy of classification can be determined using this technique, which involves resampling the data repeatedly, splitting it into training and validation folds, fitting a model with a range of model sizes on the training folds, and measuring the classification accuracy on the validation folds.

2.) **Confusion-Matrix:** A table which displays the effectiveness of ML models. There are four different projected and actual value combinations in the table: false negative, true positive, and true negative, false positive. The model's prediction of a good outcome being realized is referred by true positive. False positive refers to a model's positive prediction that is untrue. False negative refers to a model's negative prediction that is untrue. True positive denotes that a model accurately predicted a negative outcome as shown in Fig 3.2.

		------- Actual -------	
		Yes	No
Predicted	Yes	True Positives	False Positives
	No	False Negatives	True Negatives

Fig 2: Confusion matrix for yes or no variables

The confusion matrix can be used to assess recall, accuracy, and precision.

3.) **Accuracy:** The accuracy rate, also known as accuracy of classification, is determined as follows:

$$\text{Accuracy} = (Tp + Tn) / (Tp + Tn + Fp + Fn)$$

4.) **Recall:** Also known as sensitivity is the real positives that were expected to be positive is called as the Tp-rate or true positive rate. The recall was employed in this case study to rank the participants. Following is a determination:

$$\text{Recall} = Tp / (Tp + Fn)$$

The y-axis in ROC curve shows the Tp-rate.

5.) **Precision:** The actual negative rate is the genuine negatives proportions that are projected to be negative (Tn-rate) or precision. In this study, student dropout rates were evaluated with precision, and the results were as follows:

$$\text{Precision} = Tn / (Tn + Fp)$$

6.) **F1 score:** The F1score displays the harmony between two categorization measures. It is calculated as follows and reflects a trade-off computation for unbalanced data sets that is frequently used:

$$\text{F1 Score} = \sqrt{(\text{Recall} * \text{Precision})}$$

7.) **Classification Error Rate:** The classification error rate shows how many occurrences overall were incorrectly classified. The formula is as follows:

$$e = (Fp + Fn) / (Tp + Tn + Fp + Fn)$$

8.) **False Positive Rate:** The axis for x in ROC curve shows the False positive Rate. The formula is as follows:

$$\text{Fp rate} = Fp / (Fp + Tn)$$

In order to accurately identify the instances in a separate dataset and prevent over fitting, this study should take into account a number of metrics.

To run all classification algorithms 10 folds cross Validation is used. The data collection is split into ten roughly equal pieces using the 10 folds cross Validation method. The model is evaluated using the 9 remaining parts, and the validation error is calculated using the supplied part's classification. At last, the average of ten tests' findings is calculated.

5. Results and Discussion

The previous phase's models should be evaluated for their suitability for implementation, in this case study, the evaluation was based on model quantitative performance metrics rather than evaluating empirically, in which the model's creator and the subject-matter expert decide whether the model is suitable for the task at hand.

1) Comparison of the Evaluation Metrics:

For the chosen course, Table 3.1 compares the classification results from a 10 folds cross validation test using the below algorithms: LR, DT, NB, RF, NN and SVM, reporting on some of the most well-known metrics: precision, accuracy, F1 score and recall.

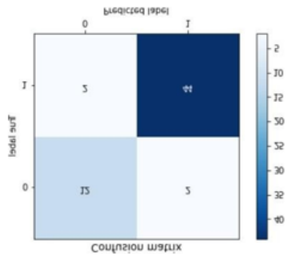
ML Algorithm	F1 score	Recall	Precision	Accuracy
NB	0.95	0.98	0.92	0.92
NN	1.00	0.93	0.97	0.93
RF	0.95	0.96	0.94	0.92
SVM	0.96	0.98	0.94	0.93
DT	0.96	0.98	0.94	0.93
LR	0.96	0.98	0.96	0.93

Table 1: Comparison of the evaluation metrics.

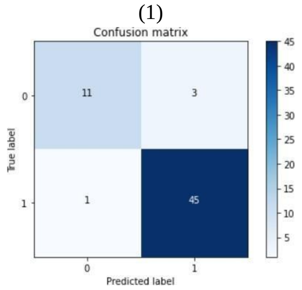
Based on the above table 1 we can say that all the 6 algorithms performed well in terms of precision, f1 score, accuracy & recall. Among these 6 the NN, SVM, DT & LR performed well with 93% accuracy.

2) Comparison of the Confusion matrices (CM):

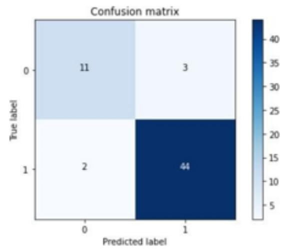
The below Fig 3.3 depicts the overall confusion matrices for the 6 different algorithms which we have utilized in our model development and are helpful in analyzing the model's performance in terms of total number of correct and incorrect predictions from the available dataset.



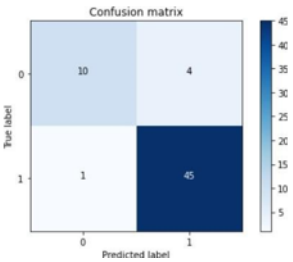
(1)



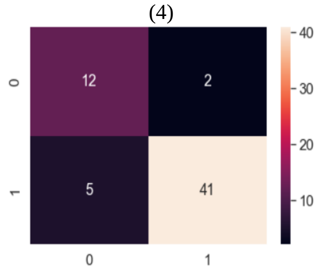
(2)



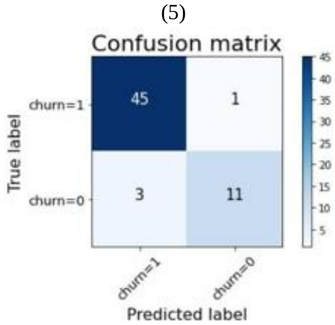
(3)



(4)



(5)



(6)

Fig 3: Confusion matrices for (1) LR, (2) DT, (3) RF, (4) NB, (5) NN, & (6) SVM.

The anticipated label is calculated from the true label in the preceding picture, where 1 – No Dropout, 0- Dropout.

From the test dataset for logistic regression, 12 cases of dropout and 44 instances of no dropout were correctly classified, whereas the other data were wrongly classified.

Utilizing the test dataset for decision tree, 11 cases of dropout and 45 instances of no dropout were correctly classified, whereas the other data were classified incorrectly.

Using the test dataset for RF, 11 dropout cases and 44 no dropout cases were successfully classified, whereas the other data were wrongly classified.

dropout cases and 45 no dropout cases were correctly classified using the test dataset for NB, whereas the other data were incorrectly classified.

Using the test dataset for NN for 100 epochs, 12 dropout cases and 41 no dropout cases were correctly identified, while the other data were wrongly classified.

dropout cases and 45 no dropout cases were successfully recognized using the test dataset for SVM, but the rest data were incorrectly classified. Below Table 3.2 indicates the overall results of the matrix of confusion for the 6 algorithms.

Table 2: Overall results of CM

Model	Total no. of correct predictions	Total no. of incorrect predictions
LR	56	4
DT	56	4
RF	56	5
NB	55	5
NN	53	7
SVM	56	4

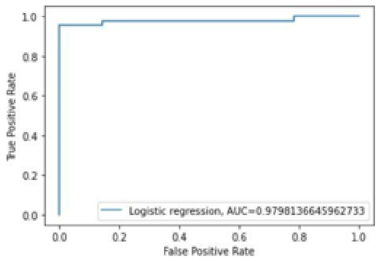
Table 3.2 interprets the CM of the overall 6 algorithms where LR, DT, SVM proves to be best in correctly predicting the dropout results.

3) Comparison of the AUC-ROC curves:

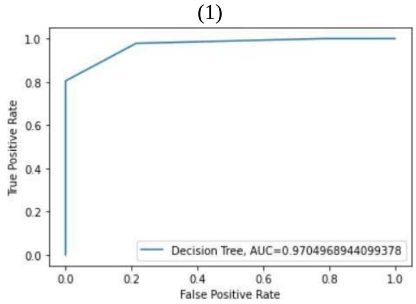
To summarize confusion matrices, AUC-ROC graphs are drawn using the Tp-rate and Fp-rate, however, there are other metrics that can do the same, like replacing Fp-rate with precision.

The efficiency of various proposed models was further examined by measuring the performance of proposed classification algorithms. To assess the effects of an unbalanced dataset, the Area under the receiver operating characteristic curve (AUC-ROC) was taken into account. The AUC- ROC is widely used to address unbalanced data since it depicts the trade-off between the false positive rate (FPR), True positive rate (TPR) in a 2 dimensional space.

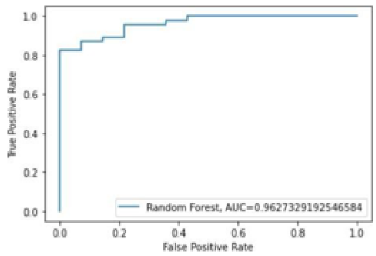
Curve of ROC is plotted from the various calculated threshold values and the TPR displays the proportion of dropouts among students that the model projected would occur in this study. The FPR, on the other hand, represents the proportion of students who did not drop out but were projected to do so by the classifier. This indicator sums up a classifier's goodness of fit.



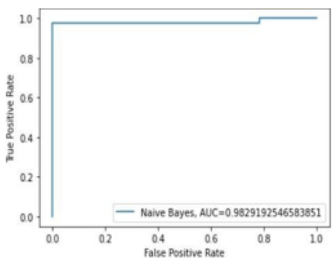
(1)



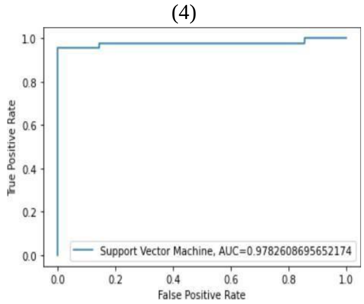
(2)



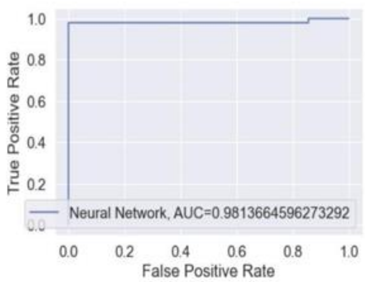
(3)



(4)



(5)



(6)

Fig 4: Comparison of AUC-ROC curves for (1) LR, (2) DT, (3) RF, (4) NB, (5) SVM and (6) NN.

As a result of our given training and testing data, the AUC-ROC curve reading for the true positive rate and false positive rate for Logistic regression is 0.9798.

Based on the dataset used for this classifier, i.e., decision tree, the AUC curve is printed with 0.9704 as the maximum.

Using the above Fig 3.4 as an example, the random forest classifier has an AUC-ROC curve value of 0.9627 for true positive and false positive rate from the test dataset.

0.9829 is the value for AUC curve for the Naïve bayes on the dataset which the algorithm has been evaluated.

The support vector machine classifier has an AUC of 0.9782 for true positive rate and false positive rate.

In the case of neural network classifiers, it is calculated that the AUC curve obtained for 100 epochs is 0.9813.

Each of the employed classification algorithms is quite effective in detecting positive class occurrences, taking into account the distribution of completers '10 i.e., class of positive and non-completers '00 i.e., class of negative from the investigated dataset. Some of them, like DT, SVM, and RF, produce a large number of false- positive predictions, which reduces the accuracy of the ultimate predictive performance. Little variations in the conditioning parameters can be distinguished by logistic regression and random forest classifier, leading to fewer failures that are predicted inaccurately. They fare better at prediction as a result.

6. Conclusion and Future Scope

On the dataset obtained from an open source, we have used six classification models in this study to predict student dropout rates: logistic regression, SVM, nave bayes, decision tree, neural networks, and RF. The findings demonstrate that none of the six models significantly differed in their ability to predict the outcomes. Random Forest, logistic regression, Decision tree and neural network classifiers are the most accurate models. We also discovered those assessments, tests, and other things affects student dropout.

In this context, it's crucial to keep in mind that identifying student who is at danger of leaving school/institutions only represents the beginning of the solution. Recognizing each student's individual needs and dropout issues is the next phase, after which adjustments are made to include successful dropout prevention techniques, which will be dealt with in our future scope where a real-time student drop out prediction model will be developed taking all these parameters into account to predict student drop out. This can illustrate the actual causes of student dropouts and offer advice on how to help students successfully complete their coursework on time.

Acknowledgements

I am thankful to Dr. Srinivas Kanakala, Assistant Professor, Department of Computer Science and Engineering at VNR Vignana Jyothi Institute of Engineering and Technology, for encouraging and guiding me with machine learning related concepts.

Author Contributions

Haarika Dasi: Conceptualization, Methodology, Software, Field study **Srinivas Kanakala:** Data curation, Writing-Original draft preparation, Software, Validation Field Study, Visualization, Investigation, Writing-Reviewing and Editing.

References

[1] Selvan, M. P., N. Navadurga, and N. Lakshmi Prasanna. "An efficient model for predicting student dropout using data mining and machine learning techniques." *International Journal of Innovative Technology and Exploring Engineering* 8.9 (2019): 750-752.

[2] Kabathova, Janka, and Martin Drlik. "Towards predicting student's dropout in university courses using different machine learning techniques." *Applied Sciences* 11.7 (2021): 3130.

[3] Sarma, C. A. ., S. . Inthiyaz, B. T. P. . Madhav, and P. S. . Lakshmi. "An Inter Digital- Poison Ivy Leaf Shaped Filtenna With Multiple Defects in Ground for S-Band Bandwidth Applications". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, Aug. 2022, pp. 55-66, doi:10.17762/ijritcc.v10i8.5668.

[4] Berens, Johannes, et al. "Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods." *Available at SSRN 3275433* (2018).

[5] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: A machine learning approach," European Journal of Higher Education, vol. 10, no. 1, p. 28–47, Jan 2020.

[6] André Sanches Fonseca Sobrinho. (2020). An Embedded Systems Remote Course. Journal of Online Engineering Education, 11(2), 01–07. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/39

[7] Yihun, Meseret. "Global challenges of students dropout: A prediction model development using machine learning algorithms on higher education datasets." *SHS Web of Conferences*. Vol. 129. 2021.

[8] P. Rawengwan and P. Seresangtakul, "A model for forecasting educational status of students," The National and International Graduate Research Conference 2017, Mar 2017.

[9] Gill, D. R. . (2022). A Study of Framework of Behavioural Driven Development: Methodologies, Advantages, and Challenges. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 09–12. https://doi.org/10.17762/ijfrcsce.v8i2.2068

[10] Fernández-García, Antonio Jesús, et al. "A real-life machine learning experience for predicting university dropout at different stages using academic data." *IEEE Access* 9 (2021): 133076-133090.

[11] A. Pradeep and J. Thomas, "Predicting college students dropout using edm techniques," International Journal of Computer Applications, vol.123, pp. 26–34, 2015.

[12] Sudais, Muhammad, et al. "Students' Academic Performance Prediction Model Using Machine Learning." (2022).

[13] Mayra, Alban, and David Mauricio. "Factors to predict dropout at the universities: A case of study in Ecuador." *2018 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2018.

[14] Prenkaj, Bardh, Giovanni Stilo, and Lorenzo Madeddu. "Challenges and solutions to the student dropout prediction problem in online courses." *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020.

[15] Sudhakar, C. V., & Reddy, G. U. . (2022). Land use Land cover change Assessment at Cement Industrial area using Landsat data-hybrid classification in part of YSR Kadapa District, Andhra Pradesh, India. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 75–86. https://doi.org/10.18201/ijisae.2022.270

[16] P. Rawengwan and P. Seresangtakul, "A model for forecasting educational status of students," The National and International Graduate Research Conference 2017, Mar 2017.

[17] Del Bonifro, Francesca, et al. "Student dropout prediction." *International Conference on Artificial Intelligence in Education*. Springer, Cham, 2020.

[18] Balraj, E., and D. Maalini. "A survey on predicting student dropout analysis using data mining algorithms." *International Journal of Pure and Applied Mathematics 118 (2018): 621-627.*

[19] W. P. Risk, G. S. Kino, and H. J. Shaw, "Fiber-optic frequency shifter using a surface acoustic wave incident at an oblique angle," *Opt. Lett.*, vol. 11, no. 2, pp. 115–117, Feb. 1986.