IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predicting Academic Performance Using Machine Learning Algorithms

Dr. S.K Singh[1], Shreya Muley[2]

[1]Head of Department (Information Technology), [2]PG Student

[1]Department of Information Technology, [2] Department of Data Science

[1,2]Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai - 400101

Abstract: In contemporary education, pupils' academic performance must be predicted correctly. Logistic Regression and Random Forest are examples of machine learning algorithms that can effectively predict students who are more likely to fail. Logistic Regression is used for the assessment and prediction of such issues as timely interventions and customized care which address multiple factors in a student's life. Conversely, Random Forest may handle big data sets well because its precision is remarkable. With a population of 480 students sampled from Kalboard360, this study compares logistic regression and random forest about demographic, educational or behavioural features. The findings reveal the effectiveness of both approaches: Logistic regression has an 81% accuracy rate with some space for improvement while Random Forest has a classification accuracy rate of 89% which shows that it can categorize different student outcomes. Such insights from these algorithms have informed tailored interventions aimed at highlighting the significance of employing Machine Learning techniques within the academic arena and understanding their strengths as well as weaknesses for successful strategies.

Keywords- Logistic Regression, Predicting Academic Performance, Educational Data Mining, Random Forest, Machine Learning

I. INTRODUCTION

The need for educators, policymakers, and researchers in today's education system to predict academic performance without any mistakes and with maximum precision is growing every day. One of the strategic approaches in this area is Predictive Analysis, which can help us identify learners who may fail to perform well academically leading them to drop out from their academic institutions. If we recognize such students ahead of time, we can do something about it by encouraging their learning process. Random Forests and Logistic Regression are identified as the two most effective methods in Predictive Analysis among others. As the frequently used statistical technique; Logistic Regression simplifies modelling binary outcomes such as whether a student will pass fail or finish the course. It considers students' historical data including previous educational achievements and social-economic background among other relevant elements that will yield predictions about the success or failure of a student. This information can be utilized by teachers when creating interventions for high-risk students and actions aimed at helping such learners because it acts as an important resource for teachers. On the other hand, Random Forest is a machine learning algorithm that employs a more complicated and flexible predictive modelling method. Random forests employ an ensemble of decision trees built on subsamples of the data to capture complex patterns and interactions among predictor variables leading to highly accurate predictions. This model is appropriate for predicting academic outcomes across diverse student populations due to its ability to handle big datasets involving multiple variables. The present study investigates the predictive accuracy of logistic regression and random forest using a large dataset containing student demographics, socioeconomic indicators, prior academic records and other relevant information. Therefore, I want to compare these two approaches to know which one has the most relative benefits for educational settings. This introductory section provides background information, the importance of this research and project objectives briefly with further discussion on the

methodology adopted in this research; findings; implications as well as recommendations for possible application in future studies or analysis thereof.

II. LITERATURE REVIEW

Recent times have seen a surge in the use of machine learning techniques to predict students' grades and identify those who may drop out. Rastrollo-Guerrero et al. (2020) emphasized the application of machine learning, collaborative filtering, recommender systems and artificial neural networks for analyzing and predicting the performance of learners [1]. Xu et al. (2019) also pointed out that usage behavior on internet is predicted using machine learning algorithms . Using machine learning techniques, first semester results can be forecasted by authors [2]. Moreover, Yağcı (2022) explored educational data mining and ML algorithms as tools to anticipate students' academic performance [3]. Furthermore, according to Xu et al. (2019), ML techniques including logistic regression classification are used for improving student's performance through identification of at-risk students as well as dropout rates. The study also found out that final grades were most accurately predicted by logistic regression classifiers among other statistical methods. Thus, this suggests that logistic regression especially is an effective tool in ML algorithm for predicting academic achievement [2]. High school dropout remains a serious problem with both personal and societal consequences. Thus, it is indispensable to detect and establish those students prone to leaving school early so as to take appropriate action in a timely manner. According to Márquez Vera et al., (2016), machine learning techniques have been increasingly used to develop early warning systems for predicting High School dropouts. In their research, Márquez Vera et al., (2016) had performed a case study about dropout prediction using machine-learning algorithms. The study showed that machine learning algorithms like artificial intelligence can be employed on high school student data sets by highlighting the ones who are likely to drop out of school. These findings demonstrate the potential of machine learning-based early warning systems in enhancing strategies for preventing dropout rates. The paper demonstrates how data-mining maybe utilized in forecasting high-school dropout rates [4]. Chung and Lee (2019) have also contributed to the literature by examining early warning systems for high school dropouts using machine learning. The source lacks sharp conclusions, but it may provide different perspectives on the efficacy of ML in predicting students' chances of dropping from schools. The present research aims at giving insight into these problems beyond what we already know [5]. Marquez et al. (2016) and Chung and Lee (2019) emphasize the rise of a machine learning interest that enables early warning systems help in identification of the dropout high school students. The focus of this review is on logistic regression and random forest applications to academic performance forecasting with respect to accuracy, precision, recall, and F1 scores. Shah et al. (2020) established the same for text classification with random forest outperforming logistic regression [6][7]. Meanwhile, Marôco et al. (2011) found an AUC of 0.865 for LOOCV –logistic regression based classifier; whereas the recall for Shah et al.'s (2020) model was 90%, with a precision of less than 61% [7][8]. This literature review indicates that logistic regression and random forest models could be used to predict academic evaluation; however it also indicates there is need for standardized reporting and further research aimed at bridging existing knowledge gaps between these models and their predictive efficacy.

III. LOGISTIC REGRESSION

Usually used in binary and multiclass classifications, Logistic Regression is a classification model that is not a regression model. It involves a dependent variable with two categorical values and one or more independent variables. Logit functions and the logistic function S are used to determine the likelihood that an observation will fit into a particular category. Using odds ratios and probability, this conclusion is also simple to reach. Because of its simplicity and readability, logistic regression is widely used in the social sciences, medical, finance, and marketing domains. I will use logistic regression to analyze precision, recall, and F1-score to distinguish between high, poor, and moderate academic performances for my project on academic performance prediction. In addition, confusion matrix analysis will be done using ROC curve analysis that can provide insight into how the model performs thus enabling early identification of problems as well as personalized learning plans aimed at enhancing students' academic performances.

IV. RANDOM FOREST

Random forests are popular for their accuracy and versatility in various applications. Several decision trees will then be developed during the training process. The outputs are computed as the averages of predictions (regression) or modes of classes (classification). One major problem with decision tree models is overfitting which can be effectively avoided by randomizing this tree creation procedure that Random Forest does. Two things are affected by randomness here, firstly, it chooses different bunches of features for

each tree, and secondly, it randomly selects data points to be used for training. It could also be said that Random Forest works well because it averages the predictions from several trees thus smoothening out any peculiarities that individual trees may contain. Random Forest also helps us to see which feature relates to what to understand the underlying dynamics in the data. Nevertheless, its heterogeneity allows it to deal with noisy and unusual datasets that interact in complicated ways since its structure is non-predictable due to resilience against anomalous events. However, compared to simpler models, random forest can have more computational complexity and still work as intended. Though this, however, was later resolved through parallel processing and optimization techniques thus making Random Forest an attractive proposition for numerous data scientists and researchers in different fields. This work will focus on implementing one aspect of the Random Forest Algorithm which is referred to as the Random Forest Classifier. In this project, the first thing that should be done is to divide the dataset into training and testing sets. Eighty per cent (80%) of the data would be used for training while twenty per cent (20%) would be reserved for testing so that it can be trained on a large amount of it. The model goes beyond accuracy only; it also considers precision, recall, and F1 scores. Also included are the ROC Curve and confusion matrix to visualize it further.

V. METHODOLOGY

5.1 Dataset

In this project, information was gathered from an LMS called Kalboard 360 and the source of this data can be located in [9]. This same information is available at Kaggle.com with the title The Students' Academic Performance Dataset. It contains 480 students' records having 16 descriptors grouped into the demographic, academic background and behavioural aspects. These are sex, nationality, educational level, grade section, parent survey responses and school satisfaction. Among them, there were 175 girls and 305 boys from different countries including Kuwait (179 students), Jordan (172 students), Palestine (28 students), Iraq (22 students), Lebanon (17 students), Tunis (12 students), Saudi Arabia (11 students), Egypt (9 students), Syria (7 students)and the United States of America, Iran and Libya (6 students from these three nations together), Morocco (4 students) and Venezuela (1 student). The database reflects a rich array of students from different backgrounds and countries. All these differences create an opportunity to deeply investigate academic achievement in diverse cultural and geographical settings. The dataset also splits students into two groups based on the number of days the student was absent; one group includes 289 students who missed less than 7 days of school while another group consists of 191 students who missed more than 7 days. Another feature category added for this dataset is Parent School Satisfaction, and Parent Answering Survey, which shows that parents play an active role in education. From the information, it can be concluded that there were 270 respondents among parents; the rest did not answer with a total number being equal to 210, and 292 parents liked their school.

VI. RESULTS

6.1 Logistic Regression

The results are therefore encouraging (Logistic regression model-based classification of the data for three categories; High (H), Low (L) and Medium (M)). A good job is done by this model in predicting the correct classes for the data samples as it has an overall accuracy of 81%. This can be seen in terms of accuracy, recall and F1 score per class. Class M which has a high recall and accuracy is also indicative that this class accurately recognizes most occurrences within its group. However, its lower recall rate makes class H not retrieve great numbers of examples from this category, though its acceptable levels of accuracy have been noticeable enough. In summary, while a logistic regression model may be able to classify data into many different groups, it seems to have the potential to recognize events classified under H only. Additionally, the Random Forest Classifier was an accurate classifier of incidents according to the confusion matrix executed in the research work. Confusion Matrix also shows the pros & cons for every class with several precise predictions. Thus we can see how ROC Curve assists in pinpointing classes themselves that's because the curve distinctly separates the classes with AUC being 0.81. This will facilitate the evaluation of how well the model has performed and determine areas that need improvement.

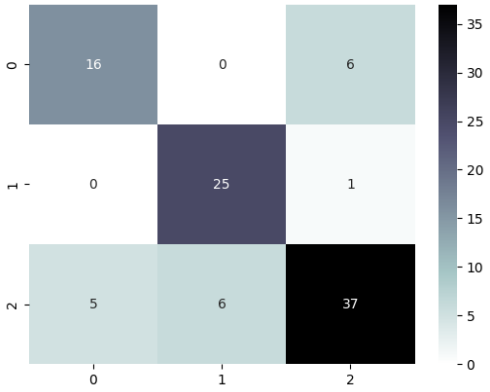

Fig.1. Confusion Matrix – Logistic Regression

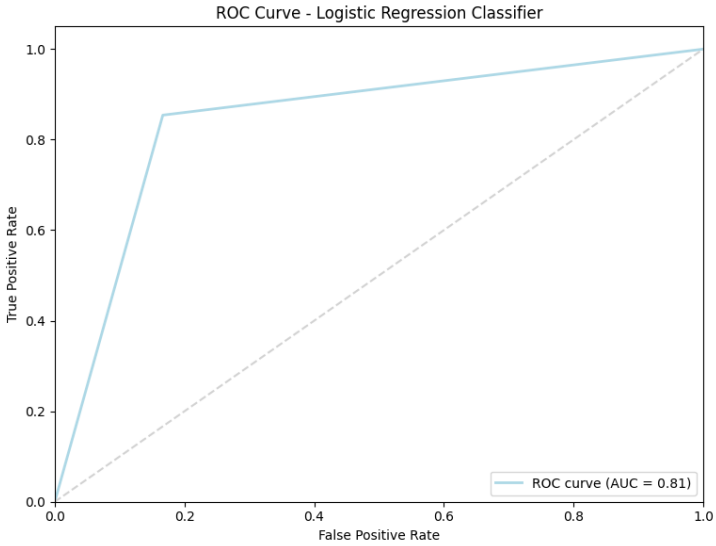

Fig.2. ROC Curve – Logistic Regression Classifier

Table 1. Results of Multiclass Model Performance for Logistic Regression Classifier

	Precision	Recall	F1 Score	Support
0	0.76	0.73	0.74	22
1	0.81	0.96	0.88	26
2	0.84	0.77	0.80	48
Accuracy			0.81	96
Macro average	0.80	0.82	0.81	96
Weighted average	0.81	0.81	0.81	96

6.2 Random Forest

By examining its examination, the Random Forest Classifier demonstrates that it can classify data into different groups. The classification report accuracy, recall and F1 score are very good predictions in many classes. How well a classification system assigns cases into appropriate categories is shown by looking at the confusion matrix. In this matrix, each cell represents the right predictions of each class which helps to determine whether a model is desirable or undesirable. Besides, the Random Forest Classifier is quite trustworthy in its forecasts as it has an overall accuracy of 89%. It tells us about different classes with high AUC in the ROC curve and 0.88 AUC means that it will correctly identify various classes. From several evaluation measures, we can say that the Random Forest Classifier does well on all fronts hence producing positive outcomes.

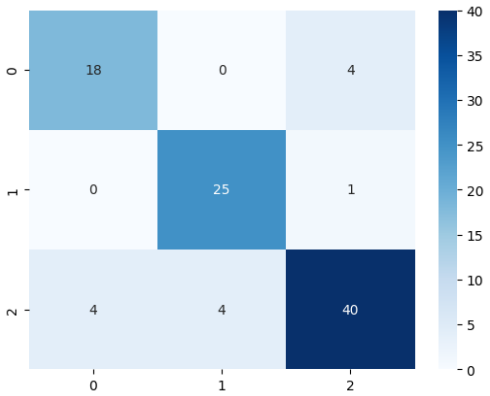

Fig.3. Confusion Matrix – Random Forest Classifier

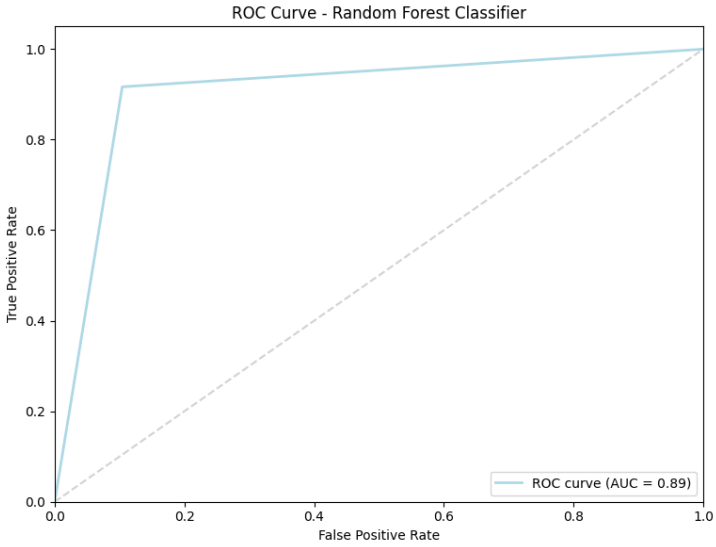

Table 2. Results of Multiclass Model Performance for Random Forest Classifier

	Precision	Recall	F1 Score	Support
0	0.82	0.82	0.82	22
1	0.86	0.96	0.91	26
2	0.89	0.83	0.86	48
Accuracy			0.86	96
Macro average	0.86	0.87	0.86	96
Weighted average	0.87	0.86	0.86	96

VII. CONCLUSION

This project, "Predicting Academic Performance using Machine Learning Algorithms" aims to use logistic regression and random forest in predicting academic performance. The results show that these methods are capable of determining student outcomes. Logistic Regression (LR) is an effective method that sorts students' education achievement into high, low and medium classes with a mean accuracy of about 81%. However, H-category discriminating cases may be improved. Random Forest Classifier has 89% overall accuracy when classifying data into multiple categories. Moreover, it can suitably adapt to complex databases making it suitable for students and data scientists alike. Both Logistic Regression and Random Forest algorithms provide invaluable insights into anticipating academic performance. To do this, one needs to know the strengths and weaknesses of different approaches to learning which will make sure that students succeed as well in using machine learning to achieve educational objectives.

VIII. REFERENCES

[1] Rastrollo-Guerrero JL, Gómez-Pulido JA, Durán-Domínguez A. Analyzing and predicting students' performance by means of machine learning: A review. Applied sciences. 2020 Feb 4;10(3):1042.

[2] Xu X, Wang J, Peng H, Wu R. Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior. 2019 Sep 1;98:166-73.

[3] Yağcı M. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments. 2022 Mar 3;9(1):11.

[4] Márquez-Vera C, Cano A, Romero C, Noaman AY, Mousa Fardoun H, Ventura S. Early dropout prediction using data mining: a case study with high school students. Expert Systems. 2016 Feb;33(1):107-24.

[5] Chung JY, Lee S. Dropout early warning systems for high school students using machine learning. Children and Youth Services Review. 2019 Jan 1;96:346-53.

[6] Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC bioinformatics. 2018 Dec;19:1-4.

[7] Shah K, Patel H, Sanghvi D, Shah M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research. 2020 Dec;5:1-6.

[8] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC research notes. 2011 Dec;4(1):1-4.

[9] Amrieh EA, Hamtini T, Aljarah I. Mining educational data to predict student's academic performance using ensemble methods. International journal of database theory and application. 2016 Aug 31;9(8):119-36.