

# Prediction of Student Academic Performance in Higher Education Institutions Using Data Mining

Kishori Kasat<sup>1</sup><sup>a</sup>, Naim Shaikh<sup>2</sup><sup>b</sup>, Meenakshi<sup>3</sup><sup>c</sup> and Khaled A. Z. Alyamani<sup>4</sup><sup>d</sup>

<sup>1</sup>*Symbiosis International (Deemed University), Pune, India*

<sup>2</sup>*Dr. D. Y. Patil Vidyapeeth (Deemed to be University), Pune, Maharashtra, India*

<sup>3</sup>*Apeejay Styu University Sohna, Haryana, India*

<sup>4</sup>*Applied College, Abqaiq Branch, King Faisal University, Saudi Arabia*

**Keywords:** Machine Learning, Random Forest, Student Performance Prediction, Accuracy, Educational Data Mining, Learning Analytics.


**Abstract:** In recent years, the analysis of student performance has emerged as one of the most significant research foci in the domains of Educational Data Mining (EDM) and Learning Analytics (LA). Because of this, a number of schools have started utilising EDM and LA to make forecasts about their students' potential for future academic success. This gives instructors and school administrators the ability to track the development of each student and take corrective action before it is too late. This article presents a model for Prediction of student academic performance in higher education institutions using data mining. This methodology section consists of data acquisition and classification phases. Data acquisition includes collecting student data on the basis of predefined attributes and classification is performed using Naïve bayes, linear regression and random forest methods. Random Forest algorithm is predicting student performance more accurately.


## 1 INTRODUCTION


There has been a huge leap forward in terms of the advancement of technology. The proliferation of such advanced technologies has led to the generation of enormous volumes of data, which are now present in every aspect of human activity. This pattern is impossible to ignore in the professional realm of academia as well (Amrieh et al, 2016). Over the course of the last several years, universities and colleges have been struggling with not one but two significant problems. To begin, there is the challenge of finding out how to utilise enormous data sets for educational purposes. This presents a number of challenges. A second significant challenge is the examination of vast volumes of educational data in order to recognise significant patterns, useful information, and correlations for the sake of educational application and decision-making. Many researchers explain that recent advances in the fields

of data mining and learning analytics have made it feasible to discover linkages and links in enormous datasets that were not before visible. In recent years, the analysis of student performance has emerged as one of the most significant research foci in the domains of Educational Data Mining (EDM) and Learning Analytics (LA) (Al-Shabandar et al, 2017). EDM components are shown in figure 1. Because of this, a number of schools have started utilising EDM and LA to make forecasts about their students' potential for future academic success. This gives instructors and school administrators the ability to track the development of each student and take corrective action before it is too late. EDM and LA have been the focus of a number of studies that have attempted to predict whether or not a student will continue their education or withdraw from it. Because accurately predicting student performance is becoming more important in the contemporary educational system, researchers are motivated to

<sup>a</sup> <https://orcid.org/0000-0003-1576-5834>

<sup>b</sup> <https://orcid.org/0000-0003-2856-0512>

<sup>c</sup> <https://orcid.org/0000-0002-4175-0508>

<sup>d</sup> <https://orcid.org/0000-0001-6894-2756>

develop modelling techniques that are both reliable and effective in their ability to do so. Technology's rapid advancement has had a significant and far-reaching effect on the realm of academic study.

Re-sampling and iterative methods are used in ML models that need a lot of computing power to increase classification accuracy. By using machine learning techniques that take subset selection into consideration, it is feasible to circumvent two issues that plague conventional classifiers, namely over-fitting and distributional demands of parameters. Emerging machine learning (ML) methods in computer science make use of logic, fundamental arithmetic, and statistics as ML techniques. However, rather than estimating the group characteristics, these methods begin with an arbitrary group separator and then tweak frequently until the classification groups are satisfied (Ajibade et al, 2019).

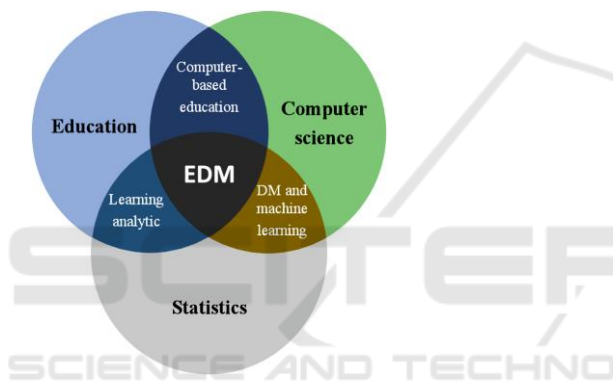


Figure 1: EDM Components.

ML does an analysis on the tuning factors as well as the ML functions that were unstable in order to locate the appropriate technique. These approaches, which are inherently non-statistical in nature, may make use of data in a wide range of formats, including nominal data, which leads to the greatest possible classification accuracies. As a consequence, these methods provide the most accurate classifications imaginable. This article presents a model for Prediction of student academic performance in higher education institutions using data mining.

## 2 LITERATURE SURVEY

Ray et al. studied educational data management and analytics (Ray et al, 2018). EDM and LA were created to help commercial and alternative communities manage massive data collections. It also describes in EDM and LA how PG shareholders' functions are affected. This document also includes

an introduction of how to use these models, assess student academic improvement, provide extensive feedback on their efforts, and analyze their academic achievement. These models influence administrative principles that work for everyone in an educational institution.

Researchers have created data preparation processes that leverage vast amounts of student data to construct marks that comply with assessment modules (Alsuaiket et al., 2020). To extract the categorical component that indicates where students' marks have been revised, data preparation must be done under various circumstances. Because of this, students' overall performance cannot be assessed without considering their specific courses. We examined the initial EDM data processing pipeline phases. Educational data differs from other types of data in origins, applications, and accuracy, hence it is generally agreed that it should be prepared differently. Thus, the coursework estimate ratio was employed to collect student transcription data to account for varied module assessment methods. A coursework assessment ratio (CAR)-based RF classifier may also improve identification.

Developers, model-relied techniques, machine learning, and data science have all noted DM as a potentially helpful collection of tools for pre-processing software development. DM is being evaluated for software expansion. Due to their focus on quantitative or qualitative research, courses that give students an introduction of this field are considered standard frameworks. This study may provide a new data science foundation for educational research (Gibson et al, 2017).

Student academic accomplishment traditionally determines a university's reputation. The popular and effective EDM prediction tool predicts students' GPA and academic performance as outstanding, very good, good, moderate, etc. It ranks pupils' academic performance as extraordinary, very good, good, average, etc. This forecasting makes it easier for institutions to select qualified scholarship recipients. Subsequently, the literature has examined the impact of subsequent features on forecasting students' academic performance or grade point average at the introductory level: authors (Sembiring et al, 2011) forecasted the final grade of 300 computer systems and software engineering students. Many multi-variate analysis methods have been used to assess a trait's importance. A supportive family can improve a child's academic performance, according to several experts. Student interest does not correlate with outcomes.

Authors studied academic performance using data from 210 undergraduate students. These criteria determine a student's grade: Research shows that considering pre-university marks and first- and second-year grades can improve graduation prediction in the final semester (Daud et al, 2017).

Researchers predicted the dropout rate of technology undergraduates using their samples (Pradeep et al, 2015). WEKA Attribute Selection Algorithms may reduce the effects of applied attributes. Post-enrollment measures like attendance, class attentiveness, and exam performance are perhaps the most crucial factors in decision-making. Higher education institutions collect vast amounts of student data, which they use in retention programs and predictive and modeling approaches to achieve effective results; age, gender, and religion are not particularly important for academic performance prediction. The college collects a student's high school transcript, SAT score, demographic data, and proof of address when they enroll. After collecting the data, you'll learn about the degree recipient's stated major, academic level, course topics, and grade. Main course modules in an online learning management system (LMS) like Moodle or BlackBoard. LMS lets users establish group discussions, view course materials, and participate in other course activities including online queries and alternative operations.

### 3 METHODS

This methodology section consists of data acquisition and classification phases. Data acquisition includes collecting student data on the basis of predefined attributes and classification is performed using Naïve Bayes, linear regression and random forest methods.

The Bayesian probability theorem is used in the Naive Bayes classifier (Sokkhey et al, 2020), which results in a straightforward and efficient approach to the classification problem. NB Classifiers are often used in the early stages of the process of text retrieval for the purpose of text categorization. The simplicity and scalability of the Naive Bayes classifier are two of its primary advantages. The NB classifier considers all of the characteristics in the dataset to be causally independent of one another, which results in an increased probability of the class variable. One is able to figure out what the conditional probability is by using Bayes' theorem. The conditional probability of an event is its likelihood given that it is connected to one or more other occurrences. The conditional probability of an event might be positive or negative. Given a hypothesis  $H$  and some evidence  $E$ , Bayesian

law states that there is a correlation between the probability of the hypothesis before receiving the evidence (denoted by the letter  $P(H)$ ) and the probability of the hypothesis after acquiring the evidence (denoted by the letter  $P(H|E)$ ).

$$P(H|E) = P(E|H).P(H) / P(E)$$

When it comes to statistics, the method of machine learning that goes by the name Logistic Regression (LR) is a well-liked option. Within the context of this model, weight features are taken from the input, logs are produced, and the data is connected linearly. LR is a statistical method that is used to solve problems involving the categorising of subjects into two groups. Classes are presumed to be almost identically different from one another. It makes use of the logistic function, which is more frequently referred to as the sigmoid function, in order to turn predictions into probabilities. Foretelling the occurrence of a binary event is accomplished via the use of a logit function (Peng et al, 2021).

RF is a kind of ensemble classifier that generates a collection of DT that may or may not be connected to one another by use randomization as the means to do so. This methodology was conceived of by Breiman and Cutler. It develops a forest of hypothetical decision-making routes by using the ensemble learning technique. Utilizing a random forest, which first generates a number of decision trees before combining them together into a single structure, may help provide accurate results for predictive modelling. The fact that a random tree may be utilised to address both classification and regression issues is the primary advantage of using its use. The mathematical expression for a random forest looks like this:  $h(y, k) = 1, 2, \dots, M$ , where  $y$  is the variable that is being input. Every DT uses a random vector as the measure, surprising quality of instances, and selects a subset of the sample data set at random to serve as the training dataset. The number of decision trees (DTs), which are represented by  $k$  in a random forest (RF) approach, the number of samples ( $n$ ) associated with each DT in the training dataset ( $M$ ), and the number of features ( $m$ ) supplied by the sample are all denoted by the variable " $m$ ." (where  $m \ll M$ ) (Dhanka et al, 2021).

### 4 RESULT ANALYSIS

The experimental inquiry takes advantage of the data gathering that was done at the university and available at UCI. This data collection has 285 different samples in total. This data collection has a total of seventeen qualities that set it apart from others

like it. Classifiers based on machine learning are first taught using 200 examples, and then they are evaluated using 85 unique cases. Results are shown in figure2, figure 3 and figure 4.

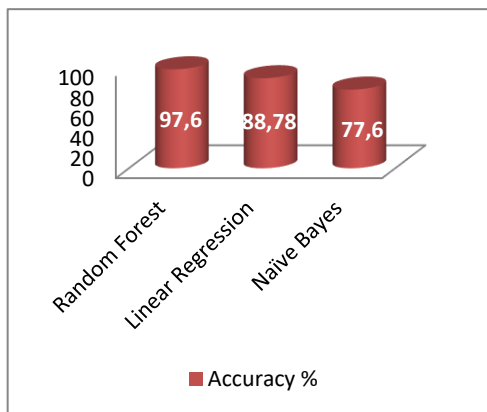


Figure 2: Accuracy of Classifiers for Student Performance Prediction.

Three metrics i.e., recall, precision and accuracy were considered. Each metric has its own view to measure the performances. Confusion matrix presents all the predicted values and hence describes the overall performance of the classification model. Recall, precision and accuracy can be calculated using the following equations:

$$Accuracy = \frac{tp_i + tn_i}{tp_i + fp_i + tn_i + fn_i}$$

$$Precision = \frac{tp_i}{tp_i + fp_i}$$

$$Recall = \frac{tp_i}{tp_i + fn_i}$$

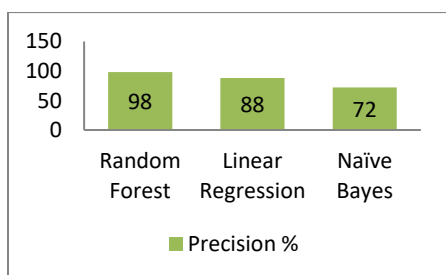


Figure 3: Precision of Classifiers for Student Performance Prediction.

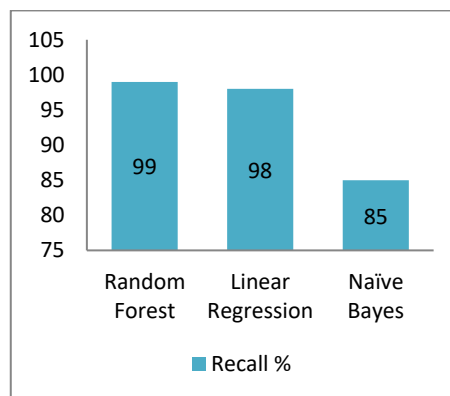


Figure 4: Recall of Classifiers for Student Performance Prediction.

## 5 CONCLUSION

In recent years, one of the most prominent research topics in the fields of Educational Data Mining (EDM) and Learning Analytics has developed to be the examination of student performance (LA). As a result of this, a number of educational institutions have begun using EDM and LA in order to create projections about the academic potential of their pupils in the years to come. This provides teachers and administrators with the capacity to monitor the progress of each kid and take remedial action before it is too late to do so. Using data mining as a model, this paper proposes a methodology for predicting the academic success of students attending higher education institutions. The steps of data collecting and categorization are included within this section on the approach. The collection of student data is accomplished using techniques such as naive bayes, linear regression, and random forest. Data acquisition also involves the gathering of student data on the basis of predetermined characteristics. The Random Forest algorithm is producing increasingly accurate forecasts of the students' performance.

## REFERENCES

- S.-S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "An heuristic feature selection algorithm to evaluate academic performance of students," 2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC), 2019. doi:10.1109/icsgrc.2019.8837067
- R. Al-Shabandar et al., "Machine learning approaches to predict learning outcomes in massive open online courses," 2017 International Joint Conference on

- Neural Networks (IJCNN), 2017. doi:10.1109/ijcnn.2017.7965922
- E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict student's academic performance using Ensemble Methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119–136, 2016. doi:10.14257/ijta.2016.9.8.13
- S. Dhanka and S. Maini, "Random Forest for heart disease detection: A classification approach," *2021 IEEE 2nd International Conference on Electrical Power and Energy Systems (ICEPES)*, 2021. doi:10.1109/icepes52894.2021.9699506
- N. S. Ruzgar and C. Chua-Chow, "Behavior of banks' stock market prices during long-term crises," *International Journal of Financial Studies*, vol. 11, no. 1, p. 31, 2023. doi:10.3390/ijfs11010031
- P. Sökkhey and T. Okazaki, "Developing web-based support systems for predicting poor-performing students using educational data mining techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020. doi:10.14569/ijacsa.2020.0110704
- A. Pradeep and J. Thomas, "Predicting college students dropout using EDM techniques," *International Journal of Computer Applications*, vol. 123, no. 5, pp. 26–34, 2015. doi:10.5120/ijca2015905328
- A. Daud et al., "Predicting student performance using Advanced Learning Analytics," *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 2017. doi:10.1145/3041021.3054164
- A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015. doi:10.1016/j.procs.2015.12.157
- D. C. Gibson and D. Ifenthaler, "Preparing the next generation of education researchers for big data in Higher Education," *Big Data and Learning Analytics in Higher Education*, pp. 29–42, 2016. doi:10.1007/978-3-319-06520-5\_4
- M. Alsuwaiket, A. H. Blasi, and R. A. Al-Msie'deen, "Formulating module assessment for improved academic performance predictability in Higher Education," *Engineering, Technology & Applied Science Research*, vol. 9, no. 3, pp. 4287–4291, 2019. doi:10.48084/etasr.2794
- S. Ray and M. Saeed, "Applications of educational data mining and learning analytics tools in handling big data in higher education," *Applications of Big Data Analytics*, pp. 135–160, 2018. doi:10.1007/978-3-319-76472-6\_7
- "University," *UCI Machine Learning Repository*, <http://archive.ics.uci.edu/ml/datasets/university> (accessed Sep. 24, 2023).