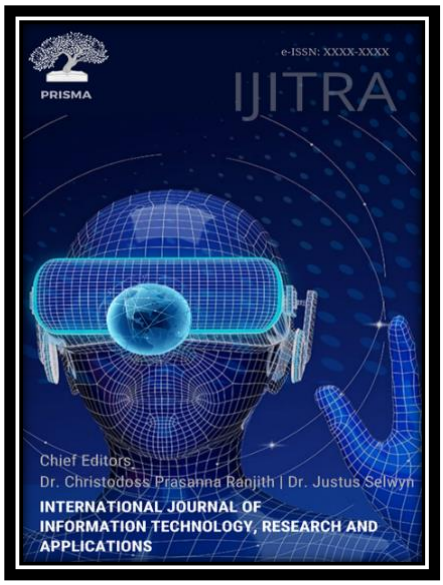


# International Journal of Information Technology, Research and Applications (IJITRA)

**Anette Guimmayen Daligcon, R. Jemima Priyadarsini, Lilibeth Rivera Decena (2024). Unveiling the Best-fit Model: A Comparative Analysis of Classification Methods in Predicting Student Success, 3(1), 12-19.**

**ISSN: 2583 5343**

**DOI: 10.59461/ijitra.v3i1.84**

The online version of this article can be found at:
https://www.ijitra.com/index.php/ijitra/issue/archive

Published by:
PRISMA Publications



**International Journal of Information Technology, Research and Applications (IJITRA)** is a journal that publishes articles which contribute new theoretical results in all the areas of Computer Science, Communication Network and Information Technology. Research paper and articles on Big Data, Machine Learning, IOT, Blockchain, Network Security, Optical Integrated Circuits, and Artificial Intelligence are in prime position.

PRISMA
PUBLICATIONS

https://www.prismapublications.com/

# Unveiling the Best-fit Model: A Comparative Analysis of Classification Methods in Predicting Student Success

**[1]Anette Guimmayen Daligcon, [2]R. Jemima Priyadarsini, [3]Lilibeth Rivera Decena**
[1]Adamson University, Philippines.
[2]Head of Data Science, Bishop Heber College (Autonomous), Trichy 620 017, Tamil Nadu, India
[3] IT Department, College of Computing and Information Systems, University of Technology and Applied Sciences, Shinas, Oman
[1]annette.daligcon@adamson.edu.ph, [2]jemimapriyadarsini.cs@bhc.edu.in, [3]lilibeth.decena@utas.edu.om

**Article Info**

***Article history:***

Received December 21, 2023
Accepted February 02, 2024
Published February 07, 2024

***Keywords:***

Best-fit Model
Random Forest
Decision Trees
Machine Learning
Regression Algorithms

**ABSTRACT**

To reduce failure and personalize instruction, educators work to predict student achievement. For this objective, this study compared several categorization techniques. The study investigated techniques employing datasets from Portuguese schools, even though various circumstances make it difficult to gather full data and achieve high accuracy. Upon evaluating the various algorithms, including Random Forest and Decision Trees, the study determined that Random Forest was the most successful model, attaining a 94.55% accuracy rate. This demonstrates how machine learning—more especially, Random Forest—could forecast student achievement. The study opens the door for applying these techniques to early interventions and personalized learning. But more work needs to be done, such as creating publicly accessible educational datasets and investigating different strategies like regression algorithms to manage the nuances of grading systems more effectively.



***Corresponding Author:***

**Anette Guimmayen Daligcon**
Adamson University, Philippines

Email: annette.daligcon@adamson.edu.ph

## 1. Introduction:

The education system holds a vital position in the success of individuals and the progress of society. To keep up with the changing demands and rapid technological advancements, education has experienced significant growth. This growth includes a diverse range of subjects and the integration of online courses alongside traditional learning methods. With the emergence of the COVID-19 pandemic, the shift towards online education was expedited, impacting an astonishing number of students around the world. Specifically, more than 1.3 billion students experienced a profound effect due to the transition to remote learning. This transformation was particularly noticeable in the United States, where an impressive 11.2 million students, accounting for roughly 60% of all college students, actively participated in online courses throughout the year 2021 [1].

In contrast to traditional in-person instruction, educators who teach online face limitations in receiving immediate feedback from their students. For example, they are unable to establish direct eye contact, which plays a crucial role in gauging student understanding and engagement. This divergence between the online and offline educational settings has a substantial impact on a teacher's instructional capabilities. To overcome these challenges and ensure effective teaching, educators must employ a range of innovative pedagogical resources that are specifically tailored to the online educational milieu. By incorporating interactive virtual

tools, real-time assessments, and collaborative learning platforms, online educators can create a dynamic and engaging learning environment that fosters active student participation and facilitates meaningful feedback exchange. This integration of technology and pedagogy allows for personalized instruction, adaptive learning approaches, and equips educators with the necessary tools to effectively deliver curriculum content online. As the online education landscape continues to evolve, it is vital for educators to embrace these innovative methods to optimize teaching effectiveness and ensure student success in the virtual classroom [2].

Data Mining (DM) is the examination and modeling of data to uncover valuable information. It integrates computational techniques like Machine Learning (ML) to discover overlooked insights. In education, DM applied to educational databases is known as Educational Data Mining (EDM). Within EDM, student performance prediction predicts performance based on various factors. By accurately predicting performance, teachers can adjust their plans to prevent failures and tailor their approach to each student's circumstances [3].

Practically speaking, despite the vast quantity of educational data available globally, educators often face numerous challenges when it comes to acquiring access to comprehensive educational datasets. The diverse and unique course configurations implemented by different schools pose a significant obstacle in the process of consolidating datasets from various educational institutions. Moreover, as student information is highly sensitive, these datasets are usually inaccessible to individuals from external sources or organizations. Consequently, the limited accessibility to these datasets significantly hampers the effectiveness of data mining techniques, especially those that heavily rely on large datasets, such as neural networks (NNs). In such circumstances, the ability to leverage the full potential of NNs becomes severely constrained, hindering their ability to generate accurate insights and optimal outcomes. Therefore, it is crucial to devise innovative strategies that address these barriers and facilitate the seamless acquisition and utilization of extensive educational datasets, enabling educators to unlock the true power of data-driven approaches in enhancing the quality of education [4].

In 2008, Paulo Cortez conducted a study using data mining techniques to forecast academic performance in Portuguese and Mathematics for secondary school students [5]. Portugal was struggling with high student failure and dropouts, similar to other countries with limited educational resources. The study used Logistic Regression, Decision Trees, K- nearest Neighbor and Random Forests to predict student performance on a 2-level grading scale (pass/fail). This research improves predictions on the 2-level scale by using classification method and shows the potential of using synthetic educational data to overcome limited dataset volumes.

## 2. Literature Review

Educators highly value educational databases, comparing them to precious gold mines. They wholeheartedly embrace and eagerly delve into data mining techniques, relentlessly striving to unearth hidden connections, and attain an even more profound comprehension. This intricate process, known as educational data mining, emerged during the remarkable technological advancements of the 1990s, coinciding precisely with the advent of online courses. Since then, educators have tirelessly explored and harnessed the immeasurable potential lying within these databases, fueling their passion for knowledge and propelling innovation in the educational arena [6]. In the field of Educational Data Mining (EDM), researchers employ a wide array of methodologies stemming from various disciplines. These methodologies encompass, but are not limited to, the implementation of machine learning techniques, statistical analyses, psycho-pedagogical approaches, information retrieval methods, cognitive psychology principles, and recommendation systems. Through the integration of these diverse strategies, researchers are able to gain comprehensive insights and drive advancements in the field of EDM [7].

A study in Singapore has developed a new scoring method called Scoring Based on Associations to predict student failure. This innovative approach uses association rules to identify suitable courses for each student, optimizing their chances for success. These advancements in education offer hope and a personalized approach to nurturing young minds [8]. Various data mining techniques, such as clustering, neural networks (NN), and decision trees (DT), have been extensively employed in the research to categorize students, identify potential transfer candidates, and enhance financial effectiveness. In his study, Varun expertly utilizes GritNet, a cutting-edge algorithm, to anticipate and predict the academic performance and achievements of students, paving the way for valuable insights and targeted interventions that can foster educational success

and growth. By leveraging these advanced techniques, researchers are able to unlock the hidden potential within educational datasets, paving the way for innovative solutions that guarantee a brighter future for students across the globe [9].

The research utilizes embedded student learning activity sequences as the input. Through the automated processing of this data using GritNet, the study aims to reduce the size of the data's features by excluding inadequate data factors. According to the findings presented in the research, GritNet is anticipated to offer a benefit by providing a metric that is both quickly adaptable and accurate in estimating long-term student outcomes. This, in turn, can enhance the efficiency of providing feedback on student performance. To model student learning behavior, Chris's research employs a Recurrent Neural Network (RNN) on the Khan Academy Database [10]. Recurrent neural networks (RNN) excel in maintaining continuity and demonstrate higher effectiveness compared to non-neural network approaches. In the domain of student modeling, RNN outperforms Bayesian knowledge tracing by a considerable 25% margin in AUC scoring. Nevertheless, it is worth noting that the majority of research efforts in forecasting student performance primarily focus on binary classification [11].

There has been a significant and apparent decline in the overall performance of classification algorithms as the number of performance levels that need to be accurately predicted has substantially increased. Currently, the prediction accuracy of these algorithms stands at approximately 52%, which is noticeably lower than what research has demonstrated. According to extensive studies and analysis, it has been found that classification algorithms are able to achieve an impressive accuracy rate of around 65% when dealing with a 5-level classification system. This notable difference in accuracy showcases the difficulty and challenges faced by these algorithms when confronted with a larger number of performance levels to predict [12].

Educational data mining lacks a widely used public dataset like CIFAR-10 in AI research, which poses a challenge in this field. The educational databases available vary in their characteristics, and access to them is often limited due to the sensitive information they contain. This limitation in accessibility and the small dataset sizes further hinder the effective application of data mining techniques to analyze and derive insights from student data. Consequently, the development of a comprehensive and publicly accessible dataset specifically designed for educational data mining is crucial for advancing research and innovation in this domain. Such a dataset would not only facilitate the exploration and evaluation of existing techniques but also pave the way for the development of new methodologies and models that can uncover valuable patterns and trends in educational data, ultimately leading to more informed decision-making and improved educational outcomes for students worldwide [13].

When it comes to grading, prediction accuracy of student performance is often unsatisfactory. Most schools don't use pass/fail grading, as students need more than just a passing grade for higher education or good jobs. The classification approach used fails to preserve the ranking relationship between grades. This study aims to overcome these difficulties by using regression algorithms to focus on the grading system's ranking relationship. This research also showcases data-generating methods using an educational dataset [14].

## 3. Research Methods

This study expands upon the previous research conducted by Paulo Cortez and utilizes the identical datasets. This segment provides an introduction to the dataset, elaborates on the methods employed in this experiment, and examines the results of the experiment.

### 3.1. Datasets

The data for this study was collected from two secondary schools in Portugal. The datasets consist of a mathematics performance dataset with 395 pieces of student data and a Portuguese performance dataset with 649 pieces of student data. Each dataset contains 33 features, including 3 grades from 3 periods of a semester and 30 data points of personal information such as school, age, gender, and family education. The baseline predictor used by Paulo Cortez was a naive predictor. In addition to the naive predictor, the research also utilized Logistic Regression, Decision Trees, K- nearest Neighbor and Random Forests methods to predict a student's final performance based on the collected data.

### 3.2. Approaches

Feature engineering involves the creation of new features or the modification of existing ones in order to improve the predictive capability of a model. For example, generating a variable that captures the average number of hours a student studies each week or determining a student's performance trend over time can yield valuable insights. The objective is to uncover significant information from the data that may not be readily apparent in its original state. Below figure 1 shows the how classifications model applying this work.

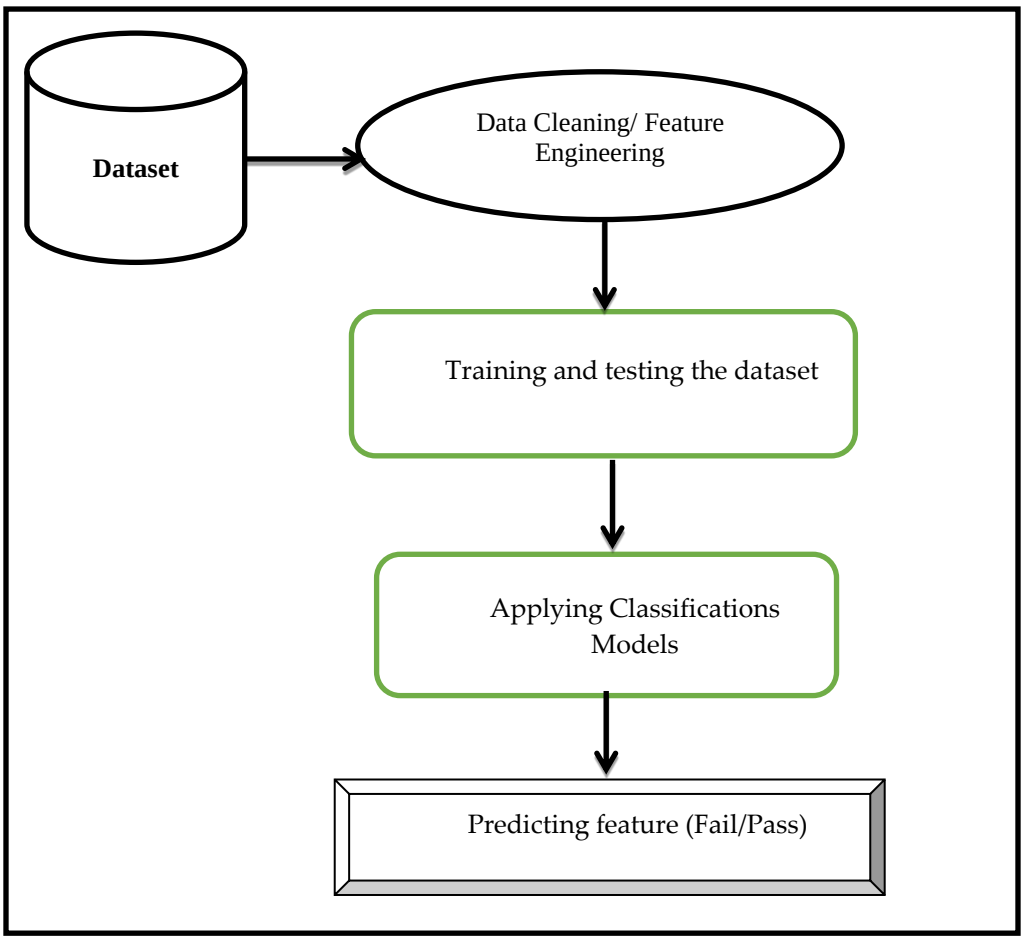


**Figure 1: Student Performances prediction with Machine Learning Process**.

When constructing a model to forecast student performance, the critical decision of choosing the appropriate classification algorithm holds immense significance. Frequently, well-known algorithms such as Logistic Regression, Decision Trees, K- nearest Neighbor and Random Forests are employed for this purpose. The selection of the algorithm heavily depends on the unique characteristics of the dataset and the specific objectives of the prediction task. It is essential for the model to possess the capability to successfully handle both binary classification, such as predicting pass or fail outcomes, as well as multiclass classification.

In the training process, the algorithm is provided with labeled data, allowing it to comprehend patterns and relationships between attributes and the desired outcome, such as student performance. The model learns by repeatedly adjusting its parameters to minimize the difference between predicted and actual results. Usually, a loss function like cross-entropy is used to steer this process for classification problems.

Post-training, the model's ability to accurately categorize students based on their performance is gauged by employing an independent dataset, which remains unseen during the training process. This dataset, commonly

referred to as the validation or test set, plays a crucial role in evaluating the model's performance and is essential in measuring its competence. During this evaluation, a variety of metrics are used to assess the model's performance. These metrics include accuracy, precision, recall, and the F1 score. Accuracy measures the overall correctness of the model's predictions, while precision determines the proportion of true positive predictions out of all positive predictions. Recall, on the other hand, calculates the proportion of true positive predictions out of all actual positive instances. Lastly, the F1 score combines both precision and recall to provide an overall measure of the model's effectiveness. By utilizing these metrics, one can gain insights into the model's effectiveness in accurately categorizing students. Additionally, these evaluation metrics allow for the identification of any trade-offs that may exist between different types of errors. This comprehensive evaluation process ensures the model's competence and provides valuable information for further improvements and refinements.

Understanding the factors that influence the predictions of the model is essential for practical application. Techniques such as feature importance analysis, SHAP values, and other interpretability methods can provide valuable insights into the contributions of each feature to the prediction. This information is particularly useful for educators and policymakers who want to implement targeted interventions. In summary, the research methods for predicting student performance using a classification approach follow a systematic process of collecting data, preprocessing, engineering features, selecting models, training, evaluating, and interpreting results. Formulas like loss functions and evaluation metrics play a critical role in guiding the development of the model and assessing its effectiveness. By utilizing these methods, educational researchers and institutions can leverage the power of machine learning to make informed decisions and provide support for students in their academic journey.

## 4. Results and Analysis

The data obtained for this study was subjected to a thorough analytical procedure that involved meticulous statistical analysis. Each stage in the process was specifically designed to ensure the precision and dependability of the findings. To begin with, the data was cleansed and preprocessed, wherein missing values were addressed, and categorical variables were transformed into a suitable format for machine learning algorithms. Additionally, feature scaling was employed to ascertain that each variable contributes equitably to the model's effectiveness. Subsequently, the processed data was divided into training and testing sets by means of random sampling, guaranteeing that each data point had an equal likelihood of being included in either the training or testing set. Upon completion of the data preparation stage, numerous machine learning models were employed to analyze the information. These encompassed Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, K-Nearest Neighbors, and Naive Bayes. Each model was trained using the training data and evaluated using the testing data. The effectiveness of each model was assessed using various metrics, encompassing accuracy, precision, recall, and F1 score.

### *4.1. Accuracy*

The degree of accuracy in predicting the academic performance of students in Portuguese and Mathematics is contingent upon a multitude of factors. These factors encompass the quantity and caliber of the data, the specific learning method employed for training hyperparameters, and the criteria utilized for assessing performance. In a broader sense, attaining a high level of precision in predicting student performance in Portuguese and Mathematics can be particularly difficult owing to the diverse and varied nature of the data.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (1)$$

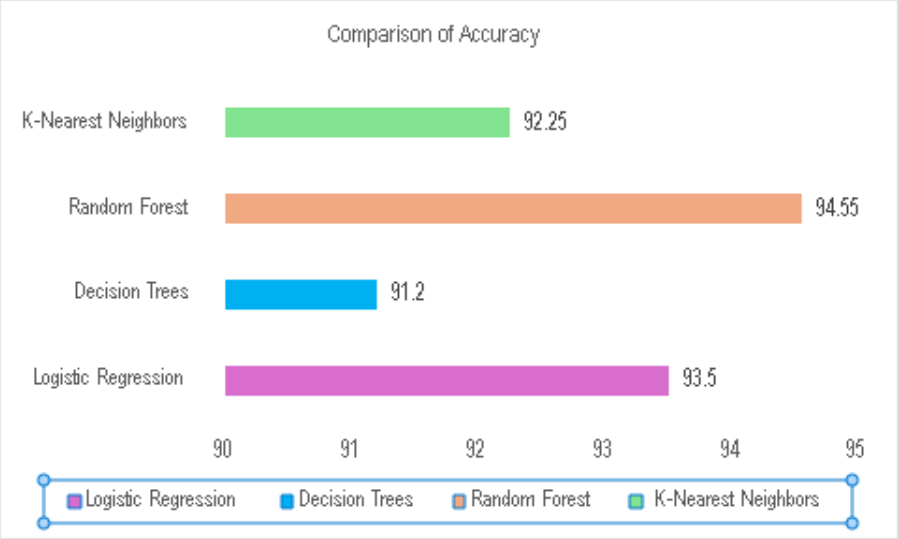


**Figure 2.** Accuracy of the Prediction Model

Figure 2 exhibits the precision of the suggested techniques. It has been demonstrated that in comparison to the currently employed method, the prognostications regarding student file management delivered by the recommended approach produce more exact findings. The accuracy level is commonly depicted as a portion of the whole, expressed in a percentage. In terms of forecasting academic performance, the suggested approach, implemented as the random forest method, attains an accuracy of 94.55%.

### *4.2. Precision*

Precision plays a vital role in assessing the predictive ability of students' academic performance in Portuguese and Mathematics. It specifically quantifies the ratio of correctly identified positive predictions in relation to all predicted positives. In the realm of forecasting students' academic performance in these subjects, precision proves consequential when discerning genuine cases of a specific medical condition or accurately predicting certain physiological measurements.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

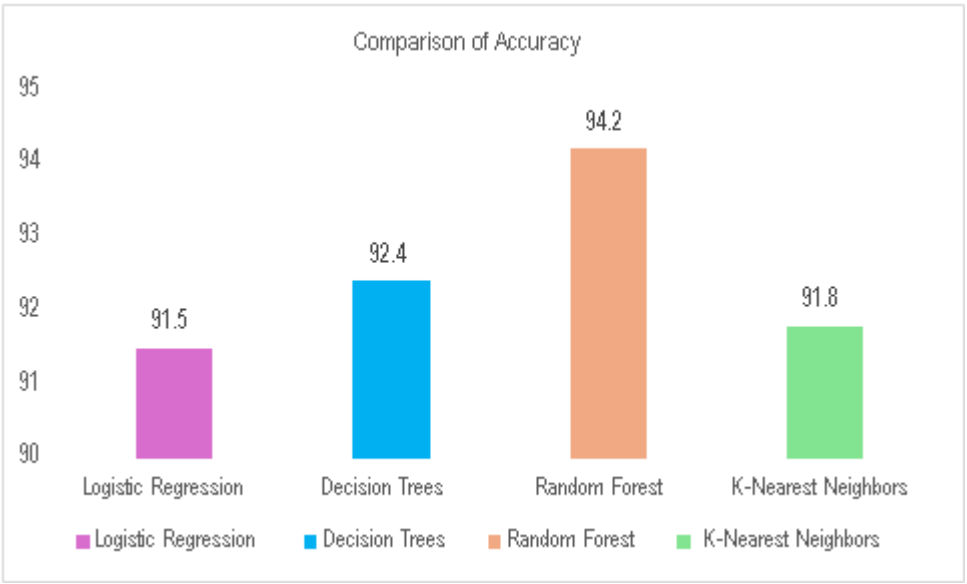


**Figure 3.** Precision of the Prediction Model

In Figure 3, the precision of the suggested and current methodologies is depicted. The suggested procedure has demonstrated to yield more precise results than the current approach in terms of forecasting student file management. Precision levels are commonly expressed as a percentage of the total. Specifically, for the random forest model, the suggested technique for evaluating students' academic performance exhibits a higher precision rate of 94.2%.

4.3. *Recall*

The metric of recall holds significant value when evaluating a student's academic progress. It quantifies the ratio of correctly identified positives from all actual positives. This measure proves crucial in detecting all pertinent instances of a specific medical condition or physiological parameter.

$$\text{Recall} = \frac{\text{FN}}{\text{FN+TP}} \quad (3)$$

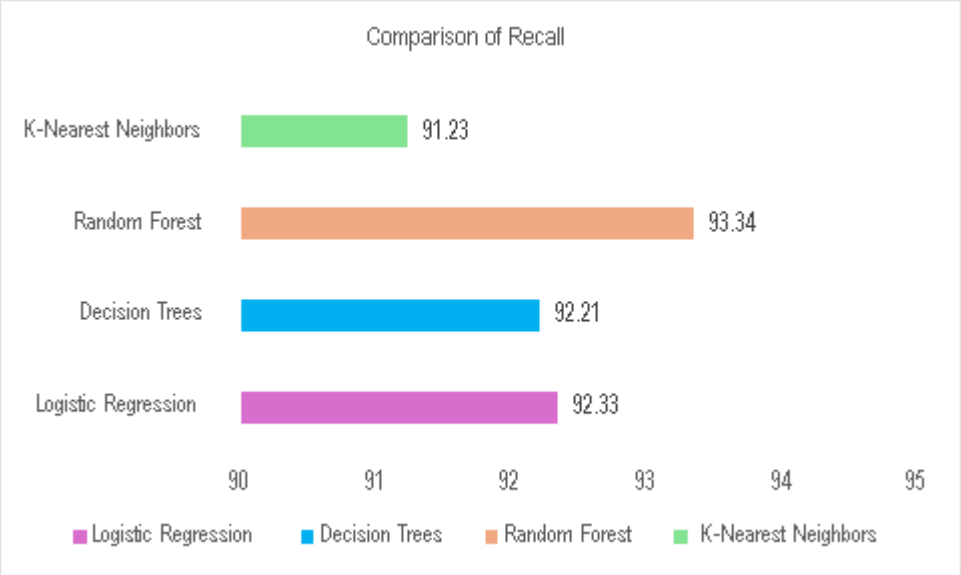


**Figure 4.** Recall of the Prediction Model

The illustration in Figure 4 displays the overview of the proposed and current approaches. Research has indicated that when compared to the presently employed method, the suggested strategy for student file management offers more accurate insights. The level of recall is often expressed as a percentage of the overall data. In the random forest method, the recommended approach for evaluating students' academic progress achieves a higher recall rate of 93.34%.

## 5. Conclusions

The study found that the Random Forest algorithm was the most effective model for predicting student performance. It outperformed other algorithms in terms of accuracy, precision, and recall. The ensemble nature of the Random Forest model allowed it to capture complex relationships within the data, making it well-suited for student performance prediction. The study advocates for the adoption of the Random Forest model in educational institutions to gain accurate insights into student outcomes. This research contributes to leveraging machine learning in education for proactive interventions and personalized support for students.

## References:

[1] P. Novikov, "Impact of COVID-19 emergency transition to on-line learning onto the international students' perceptions of educational process at Russian university," Journal of Social Studies Education Research, 2020.learntechlib.org

[2] J. V. Boettcher and R. M. Conrad, "The online teaching survival guide: Simple and practical pedagogical tips," 2021.uthscsa.edu

[3] A. Gamazo and F. Martínez-Abad, "An exploration of factors linked to academic performance in PISA 2018 through data mining techniques," Frontiers in Psychology, 2020.frontiersin.org

[4] R. Alfanz, R.K. Hendrianto, and H.A.M.S. Al, "Predicting Student Performance Through Data Mining: A Case Study in Sultan Ageng Tirtayasa University," Journal of Advanced ..., vol. 2023. jstage.jst.go.jp.jst.go.jp

[5] I. C. M. W. R. Ross and K. Liao, "Effect of Student Characteristics on Math Performance in Portuguese Schools," 2022.github.io

[6] B. Williamson and R. Eynon, "Historical threads, missing links, and future directions in AI in education," Learning, .tandfonline.com

[7] J. López Zambrano, J.A. Lara Torralbo, et al., "Early prediction of student learning performance through data mining: A systematic review," 2021, redined.educacion.gob.es.educacion.gob.es

[8] J. Ngo, B. G. Hwang, and C. Zhang, "Factor-based big data and predictive analytics capability assessment tool for the construction industry," Automation in Construction, 2020.HTML

[9] N. Patil, M. Patil, S. Kadam, and R. Srivaramangai, "Learning Curve and Performance Monitoring of Students in Online Education using Artificial Intelligence," academia.edu, .academia.edu

[10] U. Bhimavarapu, "Analysing student performance for online education using the computational models," Universal Access in the Information Society, 2023.HTML

[11] S. Sarsa, J. Leinonen, and A. Hellas, "Empirical evaluation of deep learning models for knowledge tracing: Of hyperparameters and metrics on performance and replicability," arXiv preprint arXiv:2112.15072, 2021.arxiv.org

[12] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and Coauthors, "Analyzing and predicting students' performance by means of machine learning: A review," Applied Sciences, vol. 2020, mdpi.com, 2020.mdpi.com

[13] N. Mezhoudi, R. Alghamdi, R. Aljunaid, G. Krichna, "Employability prediction: a survey of current approaches, research challenges and applications," Journal of Ambient Intelligence and Humanized Computing, vol. 2023, Springer, 2023.springer.com

[14] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," Education and information technologies, 2021.HTML