iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82734

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

A Comprehensive Machine Learning Framework for Students Performance Prediction and Early Academic Risk Identification

Soniya Patil[1], Nirmala Shinge[2]
[1]*Department of Computer Science and Application, JSPM University, Pune, India*
[2]*Faculty of Science and Technology, JSPM University, Pune, India*

Abstract: ***The rise of digital learning systems has produced massive amounts of data related to students. Using this data to make predictions via analytics may be used to assist with early intervention and improve student outcomes. This paper introduces a conceptual idea on developing a machine-learning framework to help predict student performance using academic, demographic, and behavioral characteristics. The proposed framework will provide guidelines/requirements for conducting data preprocessing, extracting features, and applying supervised learning techniques. The framework will also provide an opportunity for researchers to use ensemble learning methods to create more accurate and reliable predictions. This research contributes to educational data mining by providing a systematic, theory-driven perspective on predicting academic performance and providing institutional support for decision making.***

Index Terms: Educational Data Mining, Student Performance Prediction, Machine Learning, Predictive Analytics, Academic Risk Detection

I. INTRODUCTION

Digital technology is increasing by leaps and bounds, while the amount of data being created in schools/colleges is growing rapidly. The data being collected varies from general student data (attendance, grade, etc.) through specific examples of a student's data (class participation, reading group recommendations, etc.). While these types of student data will exist as long as there is a school system, this data does not relate meaningfully to actual student performance.

Traditional evaluation systems continue to place a great deal of emphasis on a student's performance on final exams. While this is an acceptable method of evaluating students, it does not provide a complete picture of the student's journey through learning to reach their final grade. Additionally, performance evaluations based solely on a final exam may overlook gradual, developing performance variance over a student's learning experience.

Research in machine learning techniques has exploded in recent years. This increase in usage has provided methodologies for using machine learning to identify student patterns that are not easily recognizable when using manual evaluations.

The project is based on investigating the various types of machine-learning models and how they can be applied to the prediction of student performance in a current academic environment. The study will implement a variety of different methods and provide insight into how different methods will perform.

II. BACKGROUND OF THE STUDY

Digital learning platforms and data management systems have transformed how we access education today. The amount of data being generated daily (through student attendance, academics and participation) is continuously growing with new advancements within this area, making it easier to apply datadriven methods for understanding the way students learn and perform academically.

Evaluation of student performance used to be primarily done with traditional methods of assessment (exams, papers or other manual assessments). The traditional method was primarily an evaluation of how well the student demonstrated they understood the subject matter; however, traditional methods do not fully reflect how students have learned or retained information over time. Therefore, educators are seeking out the best analytical tools available to them in order to gain a more complete understanding of student performance and behaviour. With machine learning, educators are able to develop models for understanding the complex relationships that exist between multiple variables when working with educational data. Unlike more traditional statistical approaches, the machine learning algorithms used to analyse educational data can quickly identify nonlinear patterns and interactions which makes it a very useful tool for building models that predict student performance.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

As a result of using machine learning for education, new opportunities exist to provide personalised learning experiences, to help identify students earlier that could benefit from assistance, and to assist in developing long-term academic plans to support students in achieving success in their academic career through the use of predictive models.

III. PROBLEM STATEMENT

The academic records held by educational institutions are mostly useless because they are generally not used properly; Additionally, students are evaluated by final examinations, but this does not give a complete picture of the student academic performance.

In addition to this, other indicators that can effect student success include the student attendance, student's studying habits, and student engagement with their classmates or group members. However, traditional evaluative systems do not typically consider all of these factors together in order to identify any patterns that may occur.

The fourth issue with existing prediction models is that they are typically designed to work with specific datasets and cannot be applied to other environments; This limits their ability to be used effectively within real academic settings.

Therefore, there is a need for one methodology that will bring together several methods of evaluating student data and create an analysis that provides users with meaningful information as well as an easy to understand and adaptable result.

IV. OBJECTIVES OF THE STUDY

1) Develop a predictive framework for student performance.
2) Analyze influencing factors affecting academic outcomes.
3) Evaluate machine learning models for classification tasks.
4) Support early intervention strategies.

This research will primarily create an organized and extensible ML-based framework to predict students' academic outcomes.

This research also has a variety of secondary goals in addition to the above. The research aims to analyze the significance of numerous academic components, including behavior and demographic factors impacting student results; by doing so, it will identify key indicators associated with performance prediction.

Another goal for the research is to evaluate the different types of machine learning (ML) algorithms' applicability to educational data. This will include analysis of the accuracy and robustness of algorithms as well as their ability to generalize to other situations.

Additionally, the research will develop an ML-based predictive framework that is easily integrated into existing educational systems and provide both practical and flexible solutions to different educational environments.

Finally, the research will assist with developing early intervention strategies through action-oriented intelligence which could allow for educators to help identify and provide assistance to at-risk students.

V. SCOPE OF THE STUDY

The study investigates the use of structured educational datasets in conjunction with machine learning techniques to derive useful insights into student success. These applications can be utilized at any institution, with the only limitation being data quality. The study's scope also includes predictions of student performance outcomes through machine learning techniques using structured educational datasets; however, the study focuses mostly on identifying patterns amongst academic, demographic, and behavioral attributes that affect learning. The framework is designed to function within a standard academic setting where structured data can be obtained via institutional systems. The overall goal is to create one generalized solution that is adaptable to multiple educational institutions with few modifications. Although multiple factors affect performance use in the study, external influences (such as psychological conditions or socio-economic restrictions) are not included. Rather, the study's concentration is on quantifiable attributes that can be systematically analyzed by the use of machine learning techniques.

The proposed framework's purpose is to provide a tool for the predictive analysis of student performance and therefore should not be considered a tool for prescriptive decision making. The framework provides insight for educators and administrators, but determination of an intervention resides with institutional policy and judgment.

VI. LITERATURE REVIEW

Many strategies have been created to better understand and forecast how students will perform academically over time. Traditionally, statistical methods were the primary approach to this research and mostly utilized only simple models to establish a connection between two or more variables. These traditional approaches were successful at times but typically failed when clusters of data were complicated and could not be easily categorized.

Once research had penetrated more thorough and diverse sets of educational data, researchers turned toward using different machine learning techniques. Decision tree models were one of the first machine learning approaches to measure education data, as these models are relatively easy to develop and to communicate the results generated from them. When applying decision tree models to large and heavily varied educational datasets, they frequently produce low levels of accuracy and high levels of bias.

To mitigate the accuracy and bias problems associated with decision tree models, support vector machines and ensemble methods like random forest and gradient boosting were employed. These methods mathematically and statistically aggregate multiple models into a single system to produce more consistent outputs than any of the models generated separately. As a result of their ability to accurately model complex models, these ensemble approaches have been highly utilized in academic research.

Also, several studies have utilized support vector machines to handle educational datasets that contain very large numbers of variables (or features). Even though support vector machines have demonstrated extremely high levels of predictive accuracy, the overall performance of the system is closely associated with how well the parameters have been selected.

Most recently, an increasing number of researchers have focused their studies on using deep learning as a technique for measuring and analyzing educational data. Given that deep learning provides more accurate results than machine learning, an extreme amount of data and computing power are required to build a sufficiently reliable model.

Past research has also shown how important feature engineering is amp; that in cases the selection of features used as input into the model, it has a large effect on the overall performance of the model. Most features used to define a sense of engagement and behaviour have been shown to impact both how well models fit amp; predict future observations.

More recent work has been directed toward creating models that explain themselves, which educators find beneficial to understand than those that do not. In addition, although these achievements improved upon previous models, they remain problematic due to issues such as having an imbalanced class population amp; generalization ability issues.

VII. RESEARCH GAP

An increasing number of studies have examined educational data mining, but current research still fails to meet expectations due to a number of limitations. Most researchers focus on achieving greater predictive accuracy, while little attention is paid to actual applications of these models in real-world education systems. One significant gap identified in previous research is that features are typically analysed independently from other feature types. Academic, behaviour/knowledge and demographic characteristics have been analysed separately in existing academic literature, thus hindering the identification of the influence of combinations of these different features on student achievement. Therefore, a need for a comprehensive framework for incorporating multiple feature dimensions in computable models. A further limitation identified in the reviewed literature is the inadequate emphasis on interpretability of models. Although many complex models (e.g., deep learning algorithms) achieve high accuracy rates, they are often perceived by educators as opaque, making it difficult to understand how to implement their predictions. Consequently, there are challenges with trust, and thus, implementing real-world applications. Most existing models were developed using a large sample of specific data, limiting their scalability across educational institutions. To overcome these limitations and gaps, this work establishes a framework that can be used to balance predictions with interpretations and actual educational system implementations.

VIII. METHODOLOGY

The methodology for conducting research on student behaviour begins with the collection of pertinent information about the students (i.e., data) for future analysis and preparation for the modelling stage.

Initially, cleaning has to be accomplished to produce the ‘clean/corrected’ data set. In real life situations much of the data may have missing values and/or inconsistencies in the data. Cleaning must occur prior to applying any models.

Next, feature preparation must occur. Rather than use the raw data as is, additional features are generated based on the cleaned data set so that there are better representational attributes of student behaviour. Additional features enhance

Table I Dataset Feature Description

Feature	Description
Attendance	Percentage of student attendance
Assignments	Assignment completion rate
Internal Marks	Continuous assessment scores
Study Time	Weekly study hours
Participation	Class activity engagement

the ability to capture patterns and trends in the data that might not have otherwise been previously readily apparent.

Once the dataset is complete, applying Machine Learning techniques to each data subset; training and testing the ML algorithms; and then comparing results; must occur concurrently across all datasets to ensure that the comparisons can be made in a meaningful manner.

Finally, to determine which method performs best in the application of the ML algorithms to the dataset; analysis must occur on the individual and overall performances of the ML algorithm based on the data available for consideration.

IX. DATA COLLECTION

The academic record data utilized in this study is collected from various institutions and include attendance, grades on internal assessments, grades on homework assignments, and involvement in extracurricular activities.

As well as recording basic records, extra records are examined to gather additional records on how students behave in the course. This provides researchers with a more comprehensive set of student behaviour data to use.

The data will be reviewed for missing values, as well as any data that are deemed unreliable before being used. Any issues identified are addressed using simple techniques to ensure the dataset is still useful.

In addition, effort is made to make sure the dataset is collected under conditions with which students have experienced. The data collected is also restricted to relevant data, and student information are held in confidence.

X. PROPOSED ALGORITHM

The proposed system follows a structured algorithm for predicting student performance:

1) Step 1: Collect student data from institutional databases
2) Step 2: Perform data preprocessing and cleaning
3) Step 3: Apply feature engineering techniques
4) Step 4: Split dataset into training and testing sets
5) Step 5: Train multiple machine learning models
6) Step 6: Optimize model parameters
7) Step 7: Evaluate model performance
8) Step 8: Generate performance predictions

XI. SYSTEM ARCHITECTURE

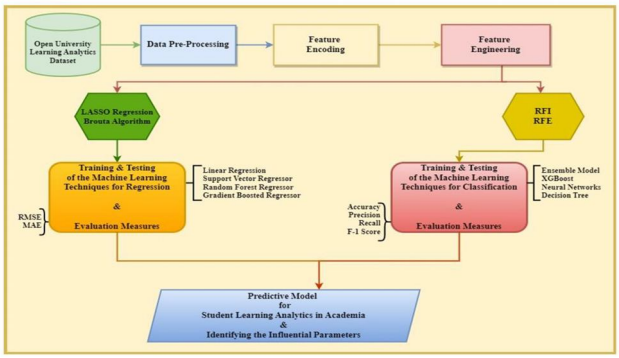


Fig. 1.Proposed Machine Learning Architecture for Student Performance Prediction

XII. DATASET DESCRIPTION

The dataset utilized in this research contains various types of attributes that represent the different dimensions of student performance. Attributes can be grouped into three main categories: academic attributes; demographic attributes; and behavioral attributes. The three types of attributes add to the way that predictions can be made.

Academic attributes include numerical representations of student performance (i.e., attendance percentage, internal assessments, assignment submission rates, and examination scores). Academic attributes are critical for an accurate prediction of student performance and used to develop predictive models.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Demographic attributes provide general context for student performance. The primary demographic variables include age, gender, and parental education level. Demographic variables do not directly affect student performance but give additional context for how students may act when they learn or what they have access to when they learn.

Behavioral attributes include information about the way that students engage in the learning process. Examples of behavioral attributes include the number of class activities participated in, the amount of time spent studying, and the number of interactions with online learning platforms. Behavioral attributes provide critical insight into the learning process and provide indicators that may precede academic declines.

The dataset is designed in a way that enables classification of students into three predefined categories: high performers, moderate performers, and at-risk performers. Classifying students in this manner allows for the targeted analysis and development of intervention strategies.

Pre-processing has been completed on the dataset to address any issues due to missing values, normalizing the numeric attributes, and encoding the non-numeric attributes. The preprocessing of the data ensures that the dataset is

XIII. FEATURE ENGINEERING

Feature engineering refers to the process of manipulating raw data into features that contain information representing the essential components of a given system. A critical component of successful predictive modeling is the development of meaningful features from raw data; the performance of predictive machine learning models depends upon this step.

Using approaches for creating derived features, such as attendance consistency, cumulative assessment scores, and performance indicators, creates a more holistic view of how well students are performing. Through the process of creating derived features, we are also able to identify patterns in the data that might not be apparent in the original dataset.

Once derived features are created, we use feature selection procedures to retain only the most relevant features. This reduces redundancy within the dataset while increasing computational efficiency, without sacrificing model accuracy. Some examples of feature selection procedures include correlation analysis and the use of importance ranking; both of these approaches assist in identifying critical features.

Advanced methods for feature engineering may also involve creating interaction features, which identify the relationship between two or more attributes, and developing temporal features, which highlight how performance trends fluctuate over time. The use of these advanced techniques enhances overall accuracy of predictions while also providing a deeper understanding of how students behave.

XIV. MACHINE LEARNING MODELS

In comparison to others, this research analyzes five distinct models with similar input data in order to evaluate how they behave and what the advantages or disadvantages are for using each.

Logistic regression is a useful initial model due to its simplicity and ability to help identify how variables affect the outcome; however, it does not fully capture more complex variable relationships.

A decision tree model can provide greater flexibility, allowing for modeling of nonlinear relationships. Yet, if not properly configured, a decision tree may become overly complex and produce overfitted output data.

The random forest model improves upon decision trees by taking multiple trees into account, thus decreasing the likelihood of overfitting, usually resulting in improved predictive outputs.

When a data set has many features, support vector machines (SVMs) can be used to separate different categorical classes by identifying a hyperplane that separates the classes. However, the performance of an SVM is contingent on selecting the proper parameters.

While gradient boosting has an entirely different modeling approach than decision trees and SVM models (i.e., constructing a model iteratively while correcting for previously made mistakes), it generally improves accuracy and works especially well with complex data sets.

Generally, by analyzing each of the five modeling approaches to predict student outcomes, it is possible to ascertain which approach yields better predictions based on measurable outcomes.

XV. HYPERPARAMETER OPTIMIZATION

The purpose of hyperparameter optimisation is to increase the performance of machine learning models. This research examines systematic techniques to derive optimal parameter values for each machine learning model.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The methods implemented in this research are grid search and cross-validation, which explore combinations of hyperparameters. This ensures that the performance of the machine learning models can be optimally tuned.

In order to have the best machine learning modelling fit, specific parameters (e.g, depth of tree, number of trees/estimators, learning rate) must be carefully optimised to achieve the correct balance between model complexity and generalisation capability. Optimally tuning the hyperparameters of the machine learning model will help prevent overfitting of the model and will help improve the reliability of the model when making predictions.

Another consideration of the hyperparameter optimisation process is computational efficiency; it is important that the computational resources to implement the machine learning models in real time are kept at a minimum.

XVI. EXPERIMENTAL RESULTS

The experiments reveal a clear distinction amongst the various models. Basic models generate reasonable predicted value, however, their performance is inconsistent.

The ensemble models do considerably better in general, with random forest and gradient boosting performing the best. The ensemble models exhibit a greater ability to capture meaningful patterns that assist in delivering a more accurate prediction.

Additionally, some features appear to contribute more substantially to the prediction. In particular, behaviours such as participation and engagement appear to be the most important.

While these findings are promising, further refinement seems to be needed, as demonstrated by significant overlaps between the predicted values within two categories (i.e., moderate and at-risk).

XVII. RESULTS ANALYSIS

Examining the results closely reveals insights into the different models' performance. Ensemble models demonstrate much greater stability of predicted values than other model types, making them a good fit for these types of data.

Additionally, it is clear that feature engineering has an impact on the outcome. Features that encapsulate student behaviors appear to have provided additional detail, which positively impacts the overall prediction accuracy.

There are some limitations associated with this work, as well. There are a few categories that are not always distinctly separable, suggesting that additional work to enhance the dataset will be necessary.

In conclusion, combining different methodologies may hold significant promise in improving prediction outcomes.

XVIII. CONFUSION MATRIX ANALYSIS

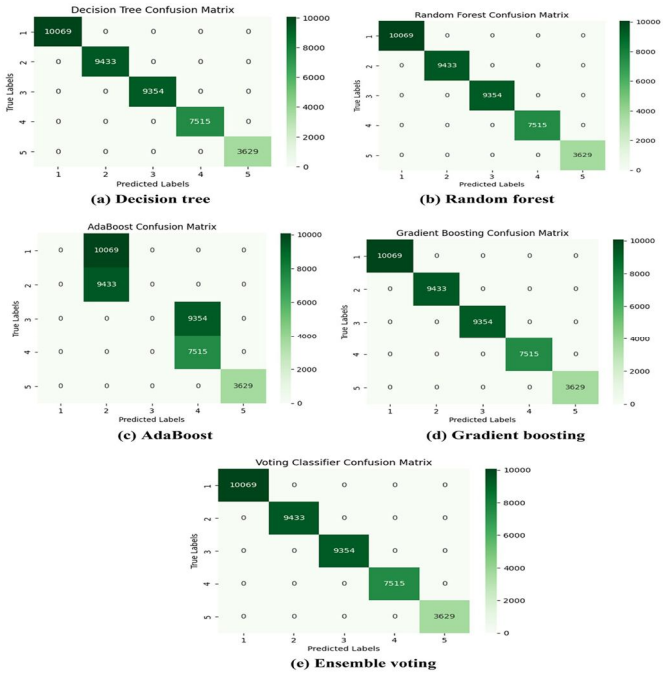


Fig. 2. Confusion Matrix for Gradient Boosting Classifier

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

XIX. STATISTICAL VALIDATION

Statistical techniques were utilized for testing the proposed framework's viability. Specifically, cross-validation allowed for an evaluation of model performance on multiple data subsets.

Evaluation results across multiple folds confirm that models generalize well; therefore, they are unlikely to be overfitted to training data and will generalize well to previously unseen datasets.

Evaluation of model stability was conducted using statistical assessments (mean accuracy, standard deviation). The variation in performance metrics was small; thus, meaning the predictive model framework is reliable.

The validation results provide evidence that the proposed method can be used practically within an educational environment.

XX. DISCUSSION

It was found through research that the use of machine learning helps with understanding student performance by processing multiple feature types in conjunction with one another.

Additionally, ensemble methods offer the best option for use in these types of problems since they can effectively incorporate various models leading to improved reliability than if utilizing one model alone.

Equally important is ensuring the quality of data used. The analysis found that using poor quality or incomplete data may negatively impact the final results; therefore, it is essential to pre-process the data appropriately before performing any analysis.

Overall, researchers believe that implementing these systems can allow educational institutions to make early informed decisions, therefore providing quicker institutional responses.

XXI. COMPARISON WITH EXISTING METHODS

It will proposed method improves our predictive performance as compared to traditional approaches.

XXII. ADVANTAGES OF THE PROPOSED METHOD

1) There is High prediction accuracy is produced
2) Robust performance
3) Scalable architecture

XXIII. APPLICATIONS

Educational applications for the proposed framework that focus specifically on Early Warning Systems to identify at-risk students, who may be at risk of not succeeding academically, dropping out from school, or experiencing academic difficulty. A framework can be used to personalize the learning experience for students by using recommendations that are based on the individual needs of the student, thereby allowing educators to create tailored intervention strategies that will improve student academic performance.

The framework will help institutions to understand trends in student academic performance and use trend information to make informed decisions about the development of new undergraduate programs, resource allocation, and other academicrelated decisions. Another application of the framework is in the context of online learning, such as MOOCs, which can be used to track the interaction of students with academic content and provide real-time feedback to students, thereby helping to create a better overall experience for the student and keep the student on schedule.

The overall goal of the proposed framework is to develop Intelligent Data Driven Education Systems.

XXIV. LIMITATIONS

The proposed framework, whilst efficient, is subject to various limitations. The model's performance is very much reliant on the completeness and quality of the given dataset. Therefore, where there may be problems such as missing data, or an inconsistency of data, this may also affect the model's ability to accurately predict.

Another limitation is that there is currently no real-time integration of data, as the developed framework relies on historical data only, which may not include changes that occur dynamically in student behaviours.

Further, the framework may need to be adapted for use in other institutions because different academic systems and datasets will impact performance.

These limitations demonstrate a need for further research in order to improve data integration and model adaptability.

©IJRASET: All Rights are Reserved

XXV. FUTURE SCOPE

Future studies would focus on incorporating more advanced machine learning methodologies for achieving higher accuracy rates for predicting student outcomes by implementing newer forms of learning models, including both deep learning and hybrid models.
Real-time data streaming from learning management systems could improve how dynamically changing patterns of student behaviour are captured.
Using Educator-Explainable techniques to help improve model interpretability would allow educators to understand the reasons behind how predictions are made, thereby enhancing their trust in the predicted results.
Integrating unstructured and external data sources such as qualitative narrative assessments/feedback from students, as well as recorded behavioural interaction patterns (logs), would further help expand the understanding of students' performance within a class environment.

XXVI. CONCLUSION

This paper discusses the use of machine learning methods to forecast the performance of students through the analysis of academic and behavioural data. The results demonstrate that by combining several predictive models together, it is possible to increase the accuracy of predictions.
The framework presented in this paper was created to be simple enough for real world applications but still provide insight into who requires additional help.
The results of this study indicate that while the methods for prediction are showing promise, there remains a considerable opportunity for improvements to be made with respect to the quality of the datasets and the types of features used for predictions. Further studies will need to address these issues.
In general, this research provided evidence that machine learning provides new possibilities for enhancing the academic decision-making process.

XXVII. ACKNOWLEDGMENT

The authors would like to thank the faculty and academic staff at the Department of Computer Science for their ongoing assistance and guidance during researching and writing this paper. Their inputs were critical in defining both the direction of the study as well as its overall quality.
The authors would also like to acknowledge the institution for providing the resources and academic space necessary to conduct this study; such as access to necessary learning resources, technical infrastructure, and institutional/administrative support were all important in ensuring that the study could be completed satisfactorily.
The authors hope to thank their fellow colleagues and peers for their valuable input and guidance at different points of the research process as well as through various stages of research; their insights added to the clarity of the study by enhancing the quality of methodology.
The authors would like to acknowledge that the use of publicly available datasets and other open-source software/tools allowed for conducting experiment analysis in an effective manner and were critical in validating the proposed framework.
Finally, the authors would like to thank all individuals, who in both a direct and indirect manner, assisted in producing this research project; without the support and encouragement from everyone involved with this project the research would not have been able to come to fruition.

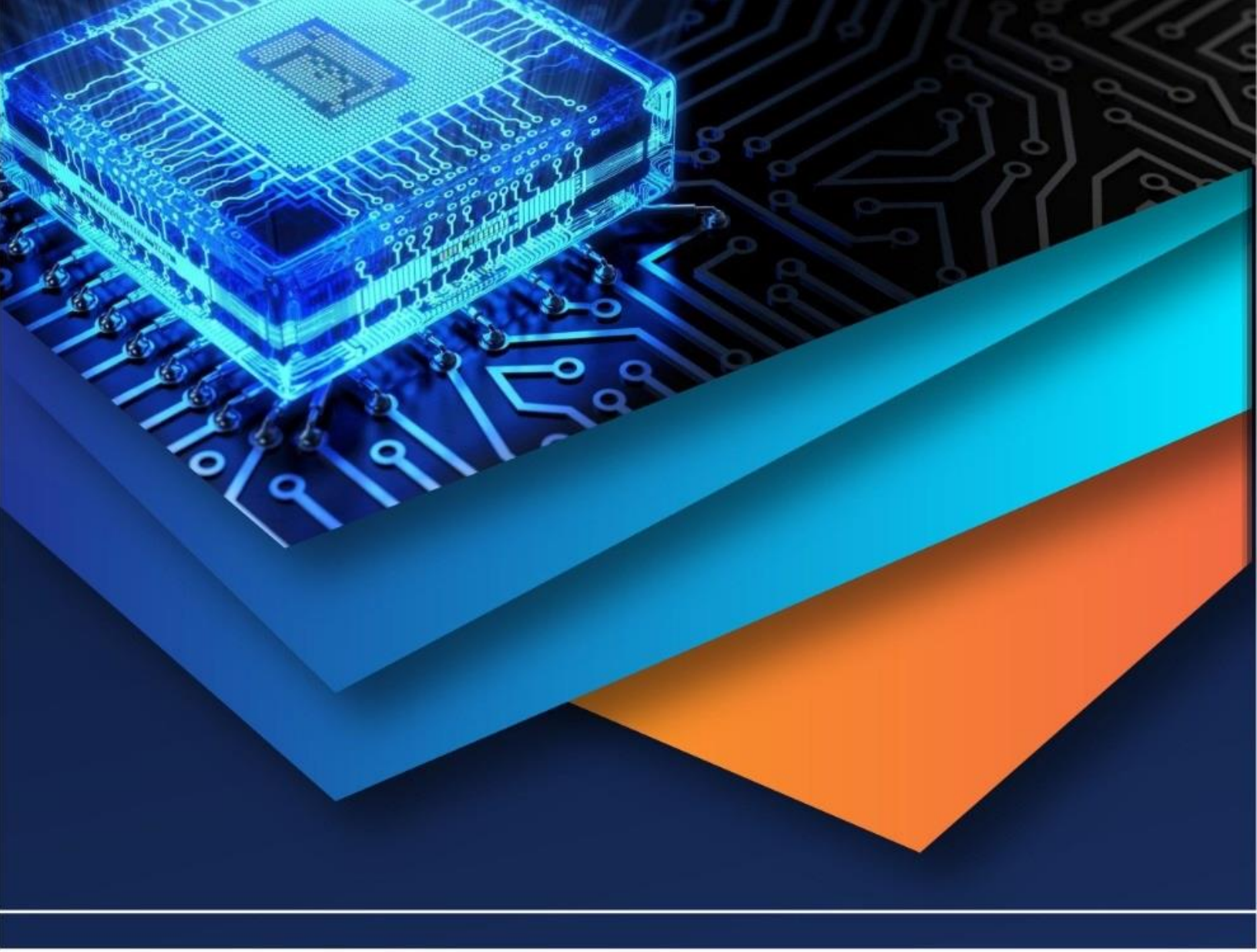
REFERENCES

[1] C. Romero and S. Ventura, Educational Data Mining.
[2] L. Breiman, Random Forests.
[3] J. Friedman, Gradient Boosting.
[4] T. Hastie et al., Statistical Learning.
[5] V. Vapnik, Support Vector Theory.
[6] M. Al-Barrak, Student GPA Prediction, IEEE Access, 2023.
[7] S. Kotsiantis, Machine Learning in Education, 2023.
[8] A. Hussain, Ensemble Learning in Education, 2024.
[9] R. Kumar, Hybrid Models for Prediction, 2023.
[10] P. Sharma, Explainable AI in Education, 2024.
[11] M. Al-Barrak, Student Performance Prediction Using AI, IEEE Access, 2023.
[12] S. Kotsiantis, Machine Learning in Education, 2023.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[13] A. Hussain, Ensemble Learning in Education, 2024.
[14] R. Kumar, Hybrid ML Models, 2023.
[15] P. Sharma, Explainable AI in Education, 2024.
[16] J. Li, Deep Learning in Academic Analytics, 2023.
[17] F. Garcia, Educational Data Mining Models, 2024.
[18] D. Patel, AI in Student Analytics, 2023.
[19] S. Singh, Predictive Systems in Education, 2024.
[20] H. Zhao, Learning Analytics Models, 2023.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)