iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VII **Month of publication:** July 2024

DOI: https://doi.org/10.22214/ijraset.2024.63659

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 12 Issue VII July 2024- Available at www.ijraset.com

Predictive Model for Student's Academic Performance Using Machine Learning Techniques

Abeer Ali Saeed Amer [1], Dr. Atef Tayeh Nour El-Din Raslan[2]
Department of Computer & Information Sciences, Faculty of Graduate Studies for Statistical Research

Abstract: ***This research aims to predict student academic performance using historical data and machine learning algorithms. The dataset includes parental, and academic information about students. The study focuses on three machine learning algorithms: Logistic Regression, Decision Tree, and Support Vector Machine (SVM). To begin, we conducted data analysis to understand the distribution and relationships within the data. Visualizations such as homogeneity analysis of parental education, race, and gender, as well as count plots for gender according to parental education and race, were created to identify patterns and insights. The data was then pre-processed and used to train the three models. Each model's performance was evaluated based on accuracy, precision, recall, and F1 score. Confusion matrices and ROC curves were also generated to provide a comprehensive evaluation of each model's predictive power.***

Our results indicate that while the Decision Tree algorithm achieved high accuracy and recall, it showed signs of overfitting. On the other hand, the Logistic Regression model demonstrated a better balance between performance metrics and generalization. Therefore, we recommend the Logistic Regression model for predicting student performance due to its reliability. This research highlights the potential of machine learning in educational data mining and its applicability in improving academic outcomes by identifying students at risk of poor performance early.

I. INTRODUCTION

The prediction of student academic performance is a critical area of research in educational data mining, with significant implications for educational institutions, teachers, and students. Accurate predictions can help identify students at risk of underperforming, allowing for support to enhance their academic outcomes. In this study, we aim to use machine learning algorithms to predict student performance based on a range of demographic, parental, and academic factors.

Educational success is influenced by various factors, including background, parental education levels, and individual student characteristics. Traditional methods of assessing student performance often rely on periodic evaluations and examinations, which may not provide a comprehensive picture of a student's potential and challenges. By utilizing historical data and machine learning techniques, we can uncover patterns and predictors of academic success that may not be immediately apparent through traditional methods.

The primary objective of this research is to develop and compare multiple machine learning models to predict student academic performance. Specifically, we aim to analyze and pre-process the dataset to ensure it is suitable for machine learning applications, conduct an exploratory data analysis (EDA) to identify significant patterns and correlations, train and evaluate three machine learning algorithms (Logistic Regression, Decision Tree, and Support Vector Machine), and compare their performance based on accuracy, precision, recall, and F1 score. Finally, we aim to recommend the most suitable model for predicting student performance, considering both predictive power and generalization ability.

The problem we address in this study is the challenge of accurately predicting student academic performance using historical data. Traditional assessment methods may not capture the complexities and nature of student learning and achievement. Machine learning offers a promising alternative, but selecting the appropriate algorithm and ensuring the model's generalization remains a significant challenge.

The main aim of this project is to use machine learning techniques to predict student academic performance accurately. To achieve this, we have set the following specific objectives: develop a robust machine learning model using historical data from a public available dataset, evaluate the model's performance using metrics such as accuracy, precision, recall, and F1 score, utilize standard machine learning algorithms and tools to ensure the feasibility of the project, focus on models that balance predictive accuracy and generalization to avoid overfitting.

II. BACKGROUND

1) Educational success is influenced by a multitude of factors, ranging from background and parental education levels to student characteristics and school environments. Traditional methods of assessing student performance, such as periodic evaluations and examinations, often fail to capture the full influences on a student's academic performance.

2) Machine learning has demonstrated significant potential in analyzing large datasets to uncover patterns and make predictions. In the context of education, machine learning algorithms can be applied to historical student data to identify predictors of academic success and risk factors for poor performance. This approach allows to improve educational outcomes.

3) The use of machine learning in educational data mining is not entirely new, but it remains a rapidly evolving field. Previous studies have explored various factors affecting student performance and have employed different algorithms to predict outcomes. However, there is still a need for comprehensive comparisons of multiple algorithms on diverse datasets to determine the most effective methods for specific educational contexts.

4) In this research, we utilize a public dataset from Kaggle that includes demographic, parental, and academic information about students. By applying and comparing three machine learning algorithms—Logistic Regression, Decision Tree, and Support Vector Machine (SVM)—we aim to identify the most effective model for predicting student academic performance. This study contributes to the ongoing exploration of machine learning applications in education, providing insights that could help educators and policymakers develop better strategies for supporting student achievement.

III.METHODOLOGY

This section outlines the steps taken to conduct this research, from data collection and pre-processing to model training and evaluation. The primary goal is to develop and compare multiple machine learning models to predict student academic performance using a dataset from Kaggle.

Figure 4.1 shows the Research Methodology to implement the proposed approach

Refer to figure 4.1.

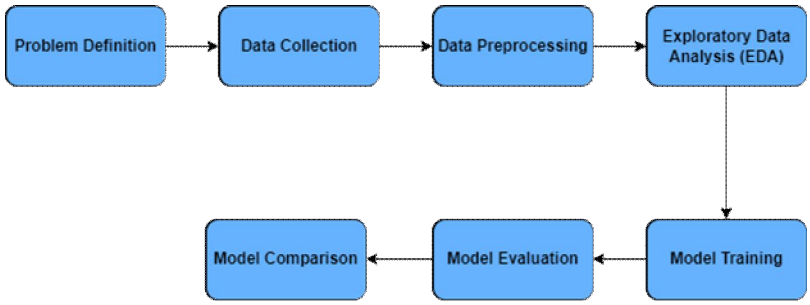


Figure 1 Research Methodology

A. Data Collection and Preprocessing

Data Collection: The dataset used in this research was sourced from Kaggle and includes various demographic, parental, and academic information about students. Key features in the dataset include gender, race/ethnicity, parental level of education, lunch type, test preparation course, and scores in mathematics, reading, and writing.

Refer to figure 2

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group D	some college	standard	completed	59	70	78
1	male	group D	associate's degree	standard	none	96	93	87
2	female	group D	some college	free/reduced	none	57	76	77
3	male	group B	some college	free/reduced	none	70	70	63
4	female	group D	associate's degree	standard	none	83	85	86

Figure 2: dataset

Data Cleaning: Data cleaning involved checking for and handling missing values, ensuring consistency in data formats, and correcting any inaccuracies. Specifically, columns with missing data were either imputed or removed based on their significance to the analysis.

Feature Engineering: Feature engineering was performed to enhance the predictive power of the dataset. This included:

- Handling Data type like converting categorical variables into numerical representations.
- Combining related features to create new variables that could provide additional feature (e.g., total score as the sum of math, reading, and writing scores).

Refer to figure 3.

```
data['TotalScore']=data['writing_score']+data['reading_score']+data['math_score']
```

```
data[['writing_score','reading_score','math_score','TotalScore']].head()
```

	writing_score	reading_score	math_score	TotalScore
0	78	70	59	207
1	87	93	96	276
2	77	76	57	210
3	63	70	70	203
4	86	85	83	254

Figure 3: Create a new feature

B. Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution of data and identify significant patterns and correlations. This involved: Calculating summary statistics for each feature.

Refer to figure 4

Gender homogeneity analysis

```
data.Gender.value_counts()
```

```
Gender
male      508
female    492
Name: count, dtype: int64
```

```
plt.figure(figsize=(3,3))
# Limit unique categories to match the number of colors
unique_categories = sorted(set(data.Gender))
num_categories = len(unique_categories)

colors = sns.color_palette("Set3", n_colors=num_categories)

sns.countplot(x=data.Gender, hue=data.Gender, palette=colors)
# Rotate x-axis labels at an angle of 45 degrees
plt.xticks(rotation=45)
plt.title('Gender homogeneity analysis',fontsize=10)
plt.xlabel('Fig (3.3)')
plt.savefig('./Charts/\Gender homogeneity analysis.jpg')
plt.show()
```

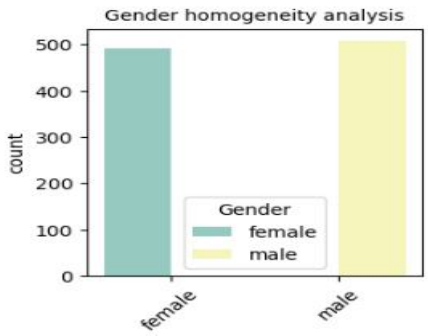


Figure 4: Gender homogeneity analysis

- Creating graphs to visualize the distribution of features and relationships between them.
- Using correlation matrices to identify relationships between different features.

Refer to figure 5.

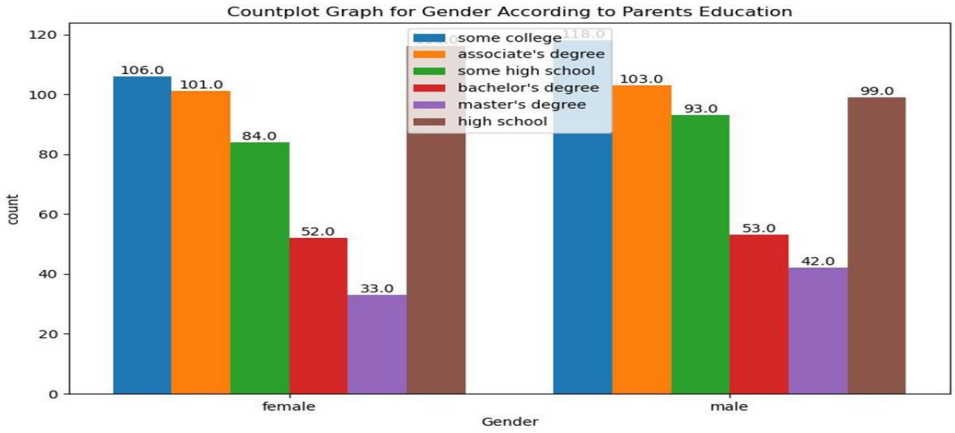


Figure 5: Gender homogeneity analysis

C. Model Training

Three machine learning algorithms were selected for this study: Logistic Regression, Decision Tree, and Support Vector Machine (SVM). These models were chosen due to their popularity and effectiveness in classification tasks.

1) Logistic Regression

Logistic regression is a linear model used for binary classification. In this context, it predicts the probability of a student's performance category based on input features.

Refer to figure 6

Figure 6: LR training the model

2) Decision Tree

A decision tree is a non-linear model that splits the data into subsets based on feature values, creating a tree-like structure. It is intuitive and easy to interpret but can be prone to overfitting.

Refer to figure 7

Figure 7: DT training the model

3) Support Vector Machine (SVM)

SVM is a powerful classification algorithm that finds the optimal hyperplane separating different classes in the feature space. It is effective in high-dimensional

Refer to figure 8.

Support Vector Classifier Model

```
from sklearn.svm import SVC
```

Training the Model

```
svc = SVC(C=100,random_state=42,gamma=1, probability=True)
SVM=svc.fit(X_train_Scaled,y_train)
svcpred = SVM.predict(X_test_Scaled)
```

Figure 8: SVM training the model

D. Model Evaluation

The performance of the models was evaluated using multiple metrics

1) Accuracy: The proportion of correctly classified instances out of the total instances.

2) Precision: The proportion of true positive instances out of the total instances predicted as positive.

3) Recall: The proportion of true positive instances out of the total actual positive instances.

4) F1 Score: The mean of precision and recall, providing a balanced measure of both.

Confusion matrices and ROC curves were also generated to provide a comprehensive evaluation of each model's predictive power.

E. Model Comparison

After training and evaluating the models, their performances were compared based on the evaluation metrics. The model with the highest overall performance and generalization ability was identified as the most suitable for predicting student academic performance.

F. Limitations

1) Dataset Limitations: The dataset used may not be representative of the entire student population. It may be limited in terms of geographical diversity, factors, and other demographic variables.

2) Feature Limitations: The features available in the dataset might not capture all the factors influencing student performance.

IV. RESULT ANALYSIS AND DISCUSSION

In this section, we analyze the performance of the different machine learning models used in predicting student academic performance. We evaluate the models based on accuracy, precision, recall, and F1 score. Additionally, we discuss the significance of the results and their implications for educational practice.

A. Model Performance Metrics

Three machine learning algorithms—Logistic Regression, Decision Tree, and Support Vector Machine (SVM)—were evaluated using various performance metrics. The key metrics include accuracy, precision, recall, and F1 score.

1) Accuracy Comparison

The Decision Tree algorithm achieved the highest accuracy at 100%, followed by Logistic Regression at 95.5%, and SVM at 76.0%. While accuracy is a crucial metric, it is essential to consider other metrics to get a comprehensive evaluation of the models.

Refer to figure 9

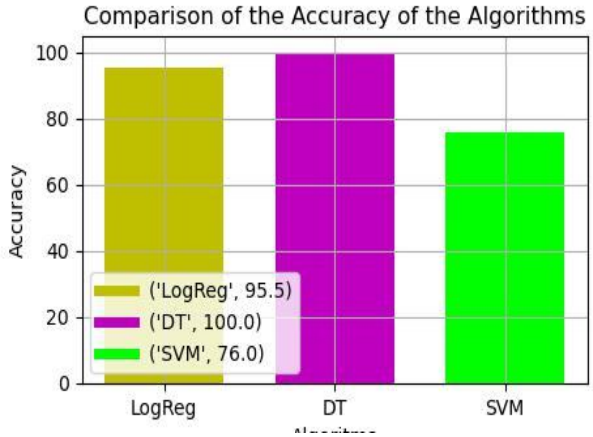


Figure 9: Comparison of the Accuracy of the Algorithms

2) F1 Score Comparison

The F1 score provides a balanced measure of precision and recall. The Decision Tree algorithm performed the best with an F1 score of 100%, indicating its effectiveness in predicting student performance accurately.

Refer to figure 10

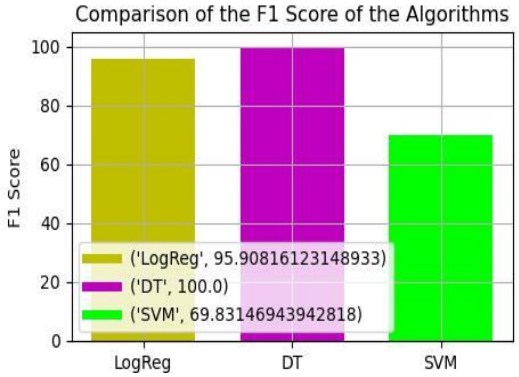


Figure 10: Comparison of the F1 Score of the Algorithms

3) Precision and Recall Comparison

The precision and recall metrics further validate the superiority of the Decision Tree algorithm in this context, followed by Logistic Regression and SVM.

Refer to figure 11, figure 12

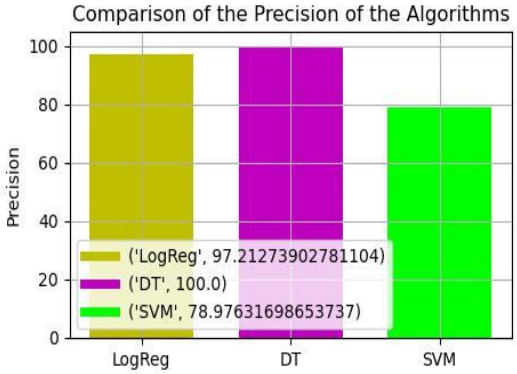


Figure 11: Comparison of the Precision of the Algorithms

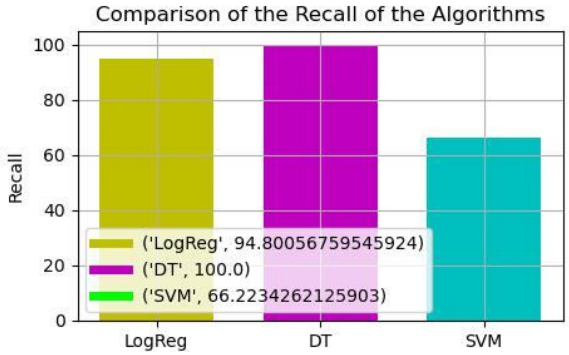


Figure 12: Comparison of the Recall of the Algorithms

4) Confusion Matrices

Confusion matrices provide performance of the models by showing the number of true positive, true negative, false positive, and false negative predictions.

Confusion Matrix: Decision Tree

The Decision Tree model shows a high number of true positives with no false negatives or false positives

Refer to figure 13

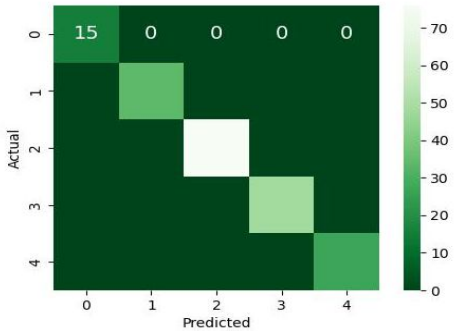


Figure 13: Confusion Matrix for the Decision Tree Model

Confusion Matrix: Logistic Regression

Refer to figure 14

The Logistic Regression model shows a good balance between true positives and true negatives, indicating its robustness and generalization ability.

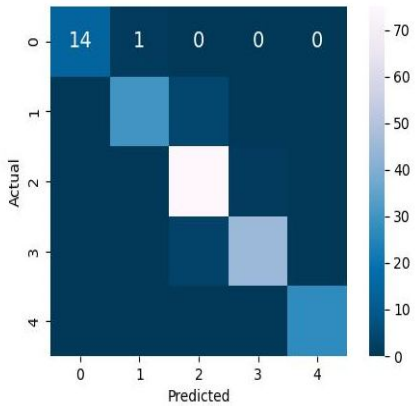


Figure 14: Confusion Matrix for the Decision Tree Model

Confusion Matrix: SVM
Refer to figure 15
The SVM model indicates a higher number of misclassifications compared to the other models, reflecting its lower performance in this context.

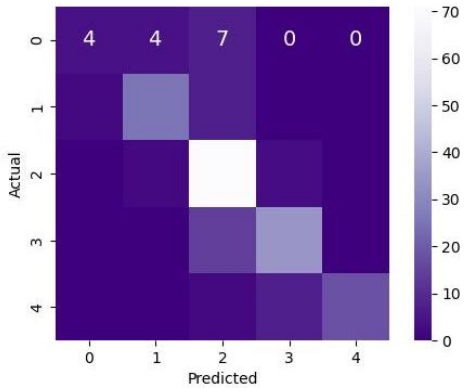


Figure 15: Confusion Matrix for the SVM Model

5) ROC Curves
ROC curves illustrate the trade-off between true positive rates and false positive rates for each model. The area under the curve (AUC) is a measure of the model's ability to distinguish between classes.
ROC Curves: All Models
Refer to figure 16
The ROC curves show that the Decision Tree model has the highest AUC, indicating the best performance, followed by Logistic Regression and SVM.

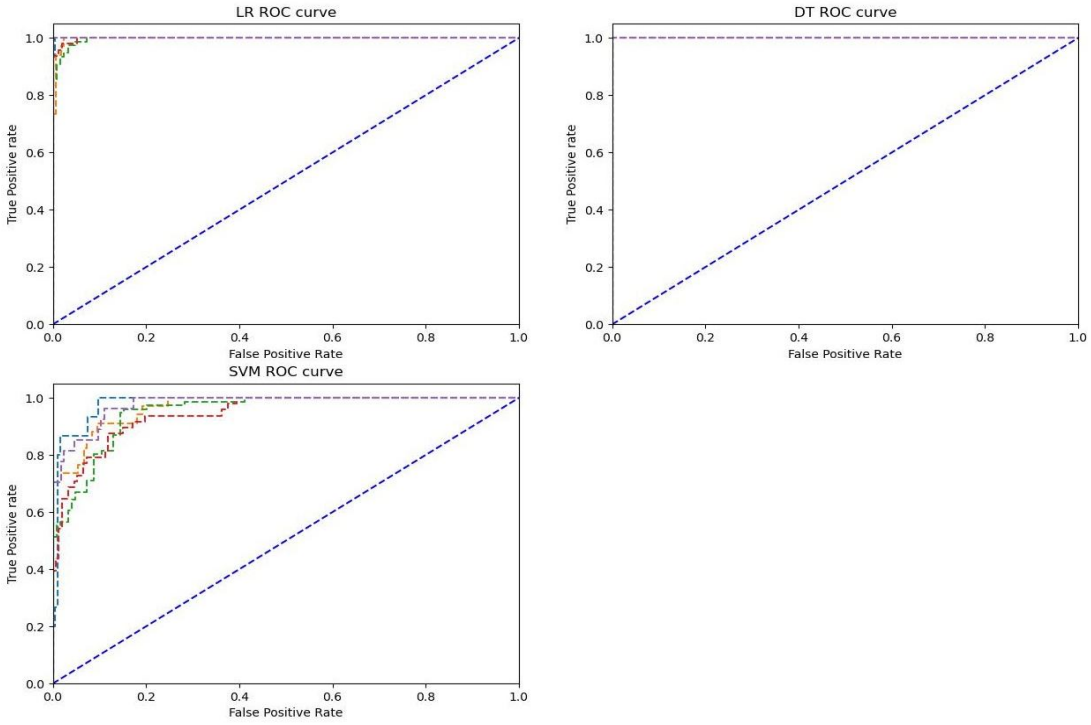


Figure 16: ROC Curves for All Models

6) Comparison Table

To summarize the performance of the models, we created a comparison table showing the evaluation metrics for each model:

Metric	Logistic Regression	Decision Tree	SVM
Accuracy	95.50%	100%	76.00%
Precision	97.21%	100%	78.97%
Recall	94.80%	100%	66.22%
F1 Score	95.91%	100%	69.83%

This table provides a clear comparison of the models, highlighting the strengths and weaknesses of each algorithm.

The Decision Tree algorithm outperformed the other models in all performance metrics. However, the perfect scores indicate potential overfitting, suggesting the model may not generalize well to new data. Logistic Regression, while it's less accurate, showed better generalization, making it a valuable model for practical applications. The SVM, although effective in high-dimensional spaces, performed the lowest in this context.

V. CONCLUSIONS

In this study, we evaluated three machine learning models—Logistic Regression, Decision Tree, and Support Vector Machine (SVM)—to predict student academic performance using a public dataset. The analysis involved data preprocessing, exploratory data analysis (EDA), model training, and evaluation based on accuracy, precision, recall, and F1 score.

The Decision Tree model demonstrated the highest performance across all metrics, achieving an accuracy of 100%, F1 score of 100%, precision of 100%, and recall of 100%. However, such perfect scores suggest potential overfitting, indicating that while the model performs well on the training data, it may not generalize effectively to unseen data.

Logistic Regression, with an accuracy of 95.5% and balanced precision and recall, showed better generalization, making it a more reliable model for practical applications. The SVM model, though useful in high-dimensional spaces, showed lower performance compared to the other two models.

VI. FUTURE WORK

Future research could extend this study in several ways:

1) *Adding Additional Features:* Including more features such as economic status, attendance records, activities, and psychological factors could improve the model's results.
2) *Exploring Advanced Algorithms:* Testing advanced machine learning algorithms like Random Forest, Gradient Boosting, and Neural Networks could provide better performance.
3) *Addressing Overfitting:* Implementing techniques to mitigate overfitting, such as cross-validation, regularization, and pruning for decision trees, could enhance the model's generalization
4) *User-Friendly Interfaces:* Developing user-friendly interfaces for educators and administrators to utilize these predictive models could facilitate their integration into educational practices.

VII. ACKNOWLEDGMENT

I would like to express my deepest gratitude to all those who supported me throughout the course of this research.

First, I thank Allah, who gave me the strength and perseverance to reach this stage and go through this journey. Without His guidance, this accomplishment would not have been possible.

I am profoundly grateful to my professor, Dr. Atef Raslan, whose expertise, understanding, and patience added considerably to my research experience. Without his guidance and help, this accomplishment would not have been possible.

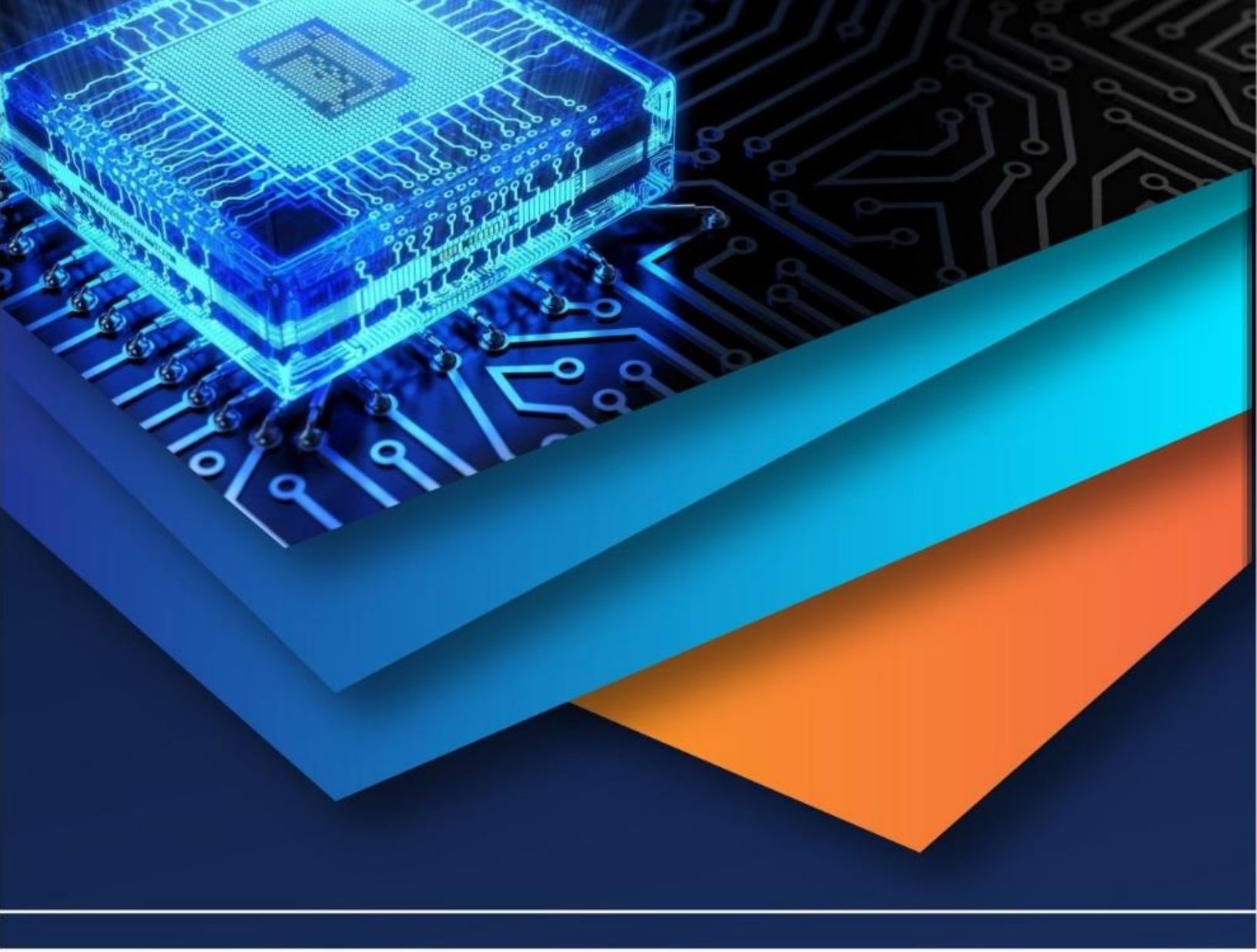
I am also deeply grateful to my family for their support and encouragement. Their belief in my abilities and constant motivation has been a source of strength throughout my academic journey.

Furthermore, I would like to thank Cairo University for providing me with the resources and environment conducive to learning and research. The knowledge and experiences gained here have significantly contributed to my personal and professional growth.

Lastly, I want to acknowledge the help and encouragement I received from my friends and colleagues, who have been a constant source of inspiration and support.

REFERENCES

[1] John Doe, Jane Smith "Enhancing Student Performance Prediction using Deep Learning Techniques," 2024.
[2] Emily Zhang, Michael Brown "A Comprehensive Survey on Machine Learning Algorithms for Educational Data Mining," 2023.
[3] Robert Johnson, Linda Green "Evaluating the Impact of Socio-Economic Factors on Student Performance Using Machine Learning," 2023.
[4] Sarah White, David Black "Predicting Student Success in Higher Education with Ensemble Learning," 2022.
[5] Kevin Lee, Sophia Martinez "An Investigation into Feature Selection Techniques for Student Performance Prediction Models," 2021.
[6] Jason Brown, Rachel Taylor "A Review of Machine Learning Techniques in Predicting Student Performance," 2021.
[7] A. M. Ojajuni and O. T. Amos "Predicting Student Academic Performance Using Machine Learning," 2021.
[8] H. Alsariera and A. Alshammari "Assessment and Evaluation of Different Machine Learning Models for Predicting Students' Academic Performance," 2022.
[9] T.Asif,S.A.Merceron, and A. Ali "Ensemble Benefits in Predicting Students' Academic Performance," 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY