ARTICLE

JOURNAL OF UNIVERSITY OF BABYLON
For Pure and Applied Sciences (JUBPAS)

Vol.33; No.1.| 2025

# Predicting Student Performance Using Data Mining and Machine Learning Techniques

**Saba Mohammed Hussain**

College of Information Technology, Dep. of Information Networks, University of Babylon, Babylon, Iraq

saba.alshebeeb@uobabylon.edu.iq

***Corresponding author email:** saba.alshebeeb@uobabylon.edu.iq mobile: 07801511455

# التنبؤ بإداء الطلاب باستخدام تقنيات التنقيب عن البيانات والتعلم الآلي

**صبا محمد حسين**

كلية تكنولوجيا المعلومات/قسم شبكات المعلومات/جامعة بابل

saba.alshebeeb@uobabylon.edu.iq

| Accepted: | 13/12/2024 | | | Published: | 31/3/2025 |
|---|---|---|---|---|---|

## ABSTRACT

Background:

Educational systems are increasingly leveraging analytic methods to improve student academic quality. Predicting student performance is an important area using data mining and machine learning can offers significant insights. This study emphasizes the potential of machine learning in environments by providing information for data-driven interventions aimed at improving student outcomes and supporting educational strategies.

Materials and Methods:

This initial processing phase guarantees model suitability. The second phase forward for training step using classifiers, like Logistic Regression Decision Tree Random Forest Support Vector Machine (SVM), K. Nearest Neighbors (KNN) and Naive Bayes.

Results:

The student results are obtained and tested using six classification models; Logistic Regression came on top with 99% accuracy; followed by SVM and K nearest Neighbors at 99%; Random Forest performed at 98%; Decision Tree, at 97%; and Naive Bayes, at 96%. The results of accuracy and recall scores showed that the models performed well, with logistic regression and k-nearest neighbors achieving around 100% perfect rates. The results of heat maps for the confusion matrix for the performance comparison reveal the effectiveness of ensemble and margin-based classifiers in this task.

Conclusions:

The results highlight how machine learning can help schools make decisions on student support and anticipate student needs effectively. For future work, the proposed work will consider metrics like behavior and engagement data and updating models consistently to improve accuracy for better student performance.

Keywords: Educational Data Mining (EDM), Decision Tree, SVM, KNN and Naive Bayes.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

## INTRODUCTION

Forecasting student performance is one of the aspects of Educational Data Mining (EDM), which may positively or negatively affect educational curricula, resource allocation, and support systems. Educational institutions aim to improve learning outcomes and identify students who may be performing poorly or failing to meet expectations, enabling timely interventions and personalized assistance. Machine learning and data analysis techniques are important tools for analyzing information and building predictive models to rank students based on their performance, identifying critical factors affecting their academic success. Data exploration is part of this field, where analytical and statistical methods and machine learning frameworks are combined to uncover valuable insights from educational datasets.

This area has received great attention in educational institutions, where huge amounts of data on student progress and performance are collected. By analyzing this data using data mining techniques, teachers can detect patterns in students learning behavior, identify areas that need improvement, and make informed decisions. Student performance research in educational data mining can also help identify students who may need support and allow interventions to improve learning outcomes and experiences [1-4].

The effectiveness of machine learning in data mining models depends on the algorithms chosen and the detailed of data preparation as shown in figure (1). Six classification algorithms; Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM) K Nearest Neighbors and Naive Bayes are implemented to predict student behavior in this work. These algorithms bring advantages when it comes to handling classification tasks. For example, Random Forest is recognized for its resilience and capability to manage data patterns that're nonlinear. Support Vector Machines deals with high dimensional environments by establishing clear boundaries between categories. Logistic Regression produces results that can be easily interpreted and is commonly employed as a benchmark in binary classification tasks. Decision Trees offer clarity in model decision making despite a tendency to overfitting which makes them valuable particularly, in educational contexts. The KNN method is non parametric, which depends on similarity measures which work well for datasets that emphasize patterns importance and in contrast NB relies on principles it is computationally efficient and appropriate for text or categorical datasets [5-8].

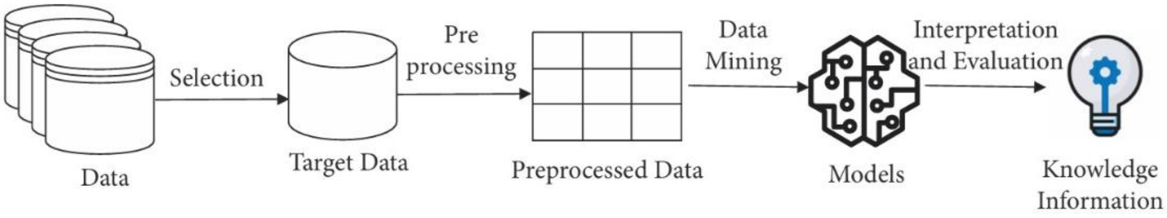

**Figure 1:** Data Mining Steps and Model Evaluation [9]

In this work, the dataset that includes information, about students' backgrounds and academic performance such as gender, parental education level, lunch type, test preparation course enrollment and scores, in math, reading and writing, is tested. The goal was to predict students' achievement levels using this data to help improve strategies. We created a performance classification variable by averaging each students scores in the three subjects (math, reading and writing). Students who scored above a threshold value were categorized as " high " while those who scored threshold were referred to as "low" using the machine learning. This distinction helps in conducting targeted evaluations. Machine learning techniques help in distinguishing between student who excel well and those who need additional guidance and assistance. In the model performance, the metrics, like accuracy along with precision and recall scores, are obtained for gauging the models results in distinguishing between low and high score accurately. Confusion matrices, for each model to show the breakdown of positives and negatives is created. Data visualization methods, alongside metrics to showcase the findings in a way that's easy to understand and interpret by creating a grouped bar graph to compare how accurate and precise each model performed by looking at their recall and F1 scores [10-11].

## RELATED WORK

In Ref. [12] in (2019) explores the application of educational data mining (EDM) techniques to improve accuracy in predicting student outcomes within a university course setting. By utilizing classification models—Naïve Bayes, Logistic Regression, k-Nearest Neighbors, and Random Forest. The study evaluated predictive models based on student engagement data from an interactive educational tool (Xorro-Q). The results indicate that the Random Forest outperformed other algorithms yielding the highest accuracy especially when combined with process mining (PM) features derived from student participation patterns. While the study demonstrated increased prediction accuracy, it acknowledged limitations related to data from a single course and lack of demographic variables. Future research could expand these findings by incorporating a broader dataset and additional student characteristics to enhance model robustness and applicability across varied educational contexts

The Ref. [13] in (2021) conducted an examination of how data mining and learning analytics are applied to forecast student performance in higher education by using machine learning methods like supervised learning and neural networks to predict grades and course completion outcomes effectively. The study classifies predictors into two categories; factors such as academic performance and nonacademic factors. One of the issues identified is the inconsistency, in models due to differences in how data gathered and the educational settings involved which make it challenging to compare studies effectively. The authors suggest that future studies should validate models using various datasets and environments while highlighting the need for interpretability to generate practical insights that can help.

In Ref. [14] (2021) investigated how Artificial Neural Networks (ANNs) are used to analyze education student information. They obtained that ANNs play a role in Educational Data Mining (EDM) helpful for predicting success and guiding personalized learning experiences for students.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل

The study highlights the significance of ANNs in developing models facilitating learning procedures and enhancing cost efficiency in handling data. Nevertheless, the research also points out difficulty in putting ideas into action, like limitations in hardware, difficulties in training and accuracy problems. After analyzing 190 articles the authors highlight the importance of developing economical methods to boost the predictive precision of artificial neural networks (ANN). Potential areas for study involve concentrating on enhancing model efficiency to raise the effectiveness of ANN in analysis personalized learning approaches and student performance which leading to sustainable progress in higher education.

IN (2022) in [9] reviewed machine learning approaches for predicting student academic performance in higher education. Their research found that ensemble methods, especially Random Forest, consistently achieved high predictive accuracy due to their ability to manage complex educational data structures. Demographic and prior academic information were significant predictors of student outcomes, and incorporating behavioral data further for improving predictions. Remarkably, the impact of admission criteria on predictive accuracy remains uncertain, suggesting a need for further study in this area. The authors highlight that interpretability and accuracy are crucial for educational models, as actionable insights can support interventions for student performance. Following Alwarthan et al., our study also evaluates multiple classifiers, including Logistic Regression and Support Vector Machine, to assess their effectiveness in real-world educational settings. Overall, these findings highlight the importance of diverse predictors and model transparency for enhancing student success .

In [15] proposed a technique to predict student exam outcomes by combining support vector machines (SVM), artificial neural networks (ANN), and teaching learning-based optimization (TLBO). This innovative method enhances the accuracy of forecasting results by selecting attributes and optimizing the ANN configuration based on the dataset from Open University encompassing engagement aspects. The results show that the Support Vector Machine (SVM), when combined with an ANN, demonstrated accuracy in classification based on input variables, while the Support Vector Regression (SVR) strength lay in regression tasks performance excellence was noted during the study into its effects, where consistent assessment scores and active participation showed a correlation with final exam outcomes this approach offers insights to improve student achievement Click or tap here to enter text..

The survey paper in [16] , there was an overview of 80 studies conducted between 2016 and 2021 in the field of educational data mining (EDM). These studies primarily revolved around forecasting student achievements through methodologies and approaches in modeling that comprise data gathering stages to assessment metrics—the phases being data collection prep work modeling development and the final evaluation process. It also categorizes different EDM techniques, compares their advantages and disadvantages, and underscores key factors such as demographic information, educational background, and levels of student engagement. In addition, the point made by the writers about the difficulties they face, like standardizing data and making predictive models more understandable due to data availability, in their studies findings emphasizes the necessity for notch standardized data and more sophisticated refining methods they propose for

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

future research directions to center on enhancing model clarity and refining features by combining insights from education and cognitive science fields for better interpretability purposes.

In Ref. [17] titled " *Identification of Students at Risk of Low Performance by Combining Rule-Based Models, Enhanced Machine Learning, and Knowledge Graph Techniques*" published in 2023, sheds light on a method to recognize and assist students who are facing challenges with their performance efficiently by integrating rule-based models with machine learning algorithms and knowledge graphs. The framework has four goals: spotting students who may need help early on in their academic journey, offering timely support by utilizing graph-based neural networks to enhance prediction accuracy, and refining personalized learning strategies based on demographic and academic data analysis using explainable machine learning techniques to better predict outcomes and provide clear explanations for the decisions made to improve student performance through customized support strategies in real-world educational settings Click or tap here to enter text..

The paper in [18] research explores a hybrid machine-learning approach to predicting student academic success. This study uses the CATboost classifier, Victoria Amazonica optimization (VAO), and artificial rabbit optimization (ARO) to enhance the accuracy of the model. By analyzing a dataset of 649 students, the researchers divided the data into training and test groups and found that the VAO model achieved an accuracy of about 6% higher than ARO. Evaluation measures, including accuracy, tuning, recall, and score F1, showed that the VAO model accurately rated 606 students, superior to other models in predictive ability. These results emphasize the potential of hybrid models in educational data mining, which offer useful tools for educators to identify at-risk students and improve academic interventions Click or tap here to enter text..

In [19] comprehensively review machine learning algorithms for predicting student success. Analyzing five algorithms—Random Forest, Support Vector Machine (SVM), Decision Tree, Logistic Regression, and K-Nearest Neighbor. The study found that Random Forest performed best, achieving a G-Mean of 0.9243 and an accuracy of 85.42%. The paper emphasizes the importance of sensitivity and balanced specificity in predictions, especially for identifying at-risk students. Key predictors included demographic and behavioral data, such as session resource sharing. The authors discuss challenges such as data standardization and model interpretability and call for further improvement to improve the applicability of machine learning in diverse educational contexts [19]

In [20] evaluated the effectiveness of four popular machine-learning algorithms: C4.5 Decision Tree, Multilayer Perceptron (MLP), Naïve Bayes (NB), and Random Forest (RF) in predicting student success. Using a dataset from secondary schools in Ghana, it was found that Naive Bayes was the best performer overall, especially in reduced datasets. The random forest also showed high accuracy, while MLP had the longest runtime and lower accuracy. The results suggest that Naive Bayes may be an ideal choice for educational data mining tasks that focus on student performance and support data-driven interventions to help at-risk students.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل
ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

## SYSTEM MODEL

The method outlined offers an organized strategy for predicting student achievements through the use of machine learning methods. It commences with data preparation wherein raw data is refined and modified to fed for machine learning models. Demographic details and other categorical attributes are encoded using Label Encoding to transform them into values that can be effectively processed by machine learning algorithms. This initial processing phase guarantees model suitability. The second phase forward for training step using classifiers, like Logistic Regression Decision Tree Random Forest Support Vector Machine (SVM) K. Nearest Neighbors (KNN) and Naive Bayes. Decision Trees captures relationships in a way that's easy for humans to interpret. Random Forests enhance accuracy and reliability by combining decision trees. SVM works well with data that has distinct class separations. KNN uses similarity measures for grouping data into classes. Naive Bayes offers efficiency with its assumptions. Using a variety of algorithms in this approach able to evaluate model's capabilities in predicting student performance[21-22] .

The training process for each model is assessed using test dataset to measure how well it predicts student outcomes accurately and reliably. For an evaluation of each classifier's performance accuracy, precision, recall, and F-scores are calculated from the confusion matrix. Accuracy gives an indication of the classifier's correctness in its predictions whereas precision and recall provide an understanding of how the model can correctly identify high performing students. The f-score balances between precision and recall by encompassing both positives and false negatives in a metric. These measurements provide an insight into the strengths and weaknesses of each classifier to guarantee that the chosen model excels in accurately investigating students who may be low or high achievement. The main methodology steps are illustrated in figure (2).

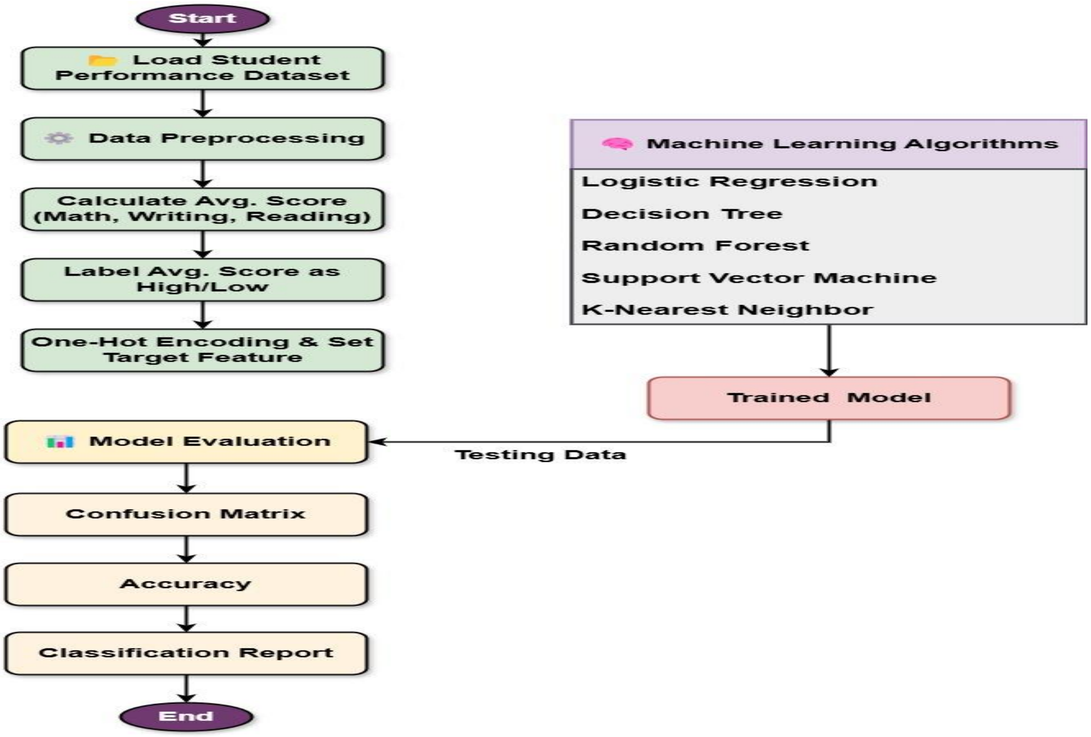

**Figure 2:** The main steps of the Student Performance Methodology

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

The pseudo code is illustrated in algorithm (1) to predict student performance.

**Algorithm 1 : Pseudocode for Predicting Student Performance Using Machine Learning**

```
BEGIN
# Step 1: Load Data
    LOAD data from 'StudentsPerformance.csv'
# Step 2: Data Preprocessing
    CALCULATE 'average_score' as mean of 'math score', 'reading score', 'writing score'
    DEFINE 'performance' based on 'average_score' (IF average_score >= 70 THEN 'high'
    ELSE 'low')
    ENCODE categorical features (e.g., gender, parental level of education)
# Step 3: Split Data
    SET X as features (all columns except 'performance')
    SET y as target (column 'performance')
    SPLIT data into training set (X_train, y_train) and test set (X_test, y_test) with 80-20
    ratio
# Step 4: Initialize Classifiers
INITIALIZE classifiers:
    Logistic Regression
    Decision Tree
    Random Forest
    Support Vector Machine (SVM)
    K-Nearest Neighbors (KNN)
    Naive Bayes
# Step 5: Train, Predict, and Evaluate Each Classifier
FOR each classifier in classifiers:
    TRAIN classifier on (X_train, y_train)
    PREDICT y_pred using classifier on X_test
  # Calculate Evaluation Metrics
    accuracy = CALCULATE accuracy_score(y_test, y_pred)
    precision = CALCULATE precision_score(y_test, y_pred)
    recall = CALCULATE recall_score(y_test, y_pred)
    f1_score = CALCULATE f1_score(y_test, y_pred)
    PRINT classifier name and its metrics (accuracy, precision, recall, F1 score)
  # Confusion Matrix Visualization
    GENERATE confusion matrix for y_test vs y_pred
    DISPLAY confusion matrix heatmap
END FOR
# Step 6: Results Visualization
    CREATE bar chart for metrics (accuracy, precision, recall, F1 score) across all classifiers
    DISPLAY bar chart
END
```

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

## SIMULATION RESULTS AND DISCUSSION

The study results offer an analysis of the performance for six machine learning models. Logistic Regression Decision Trees, Random Forests, Support Vector Machines (SVM) K. Nearest Neighbors (KNN) and Naive Bayes. The dataset was download from Kaggle website that made marks secure by different subjects. The dataset consists of many attributes as in table (1) for 1000 student [23].

Table 1: The main attributes of Student performance dataset

| No. | Attribute Name | Description |
|---|---|---|
| 1 | Gender | male/female |
| 2 | race/ethnicity | group A, B, C, D, and E |
| 3 | parental level of education | High school, some college, bachelor's degree, , master's degree, associate's degree, |
| 4 | Lunch | Free/ reduce or standard |
| 5 | test preparation course | Completed / none |
| 6 | Scores | Math/ reading / writing |

Table (2) shows the snapshot of the original dataset for the student performance which consist of 1000 rows and 8 columns.

**Table (2): Example of Student Performance Dataset**

| gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score |
|---|---|---|---|---|---|---|---|
| female | group B | bachelor's degree | standard | none | 72 | 72 | 74 |
| female | group C | some college | standard | completed | 69 | 90 | 88 |
| female | group B | master's degree | standard | none | 90 | 95 | 93 |
| male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 |
| male | group C | some college | standard | none | 76 | 78 | 75 |
| female | group B | associate's degree | standard | none | 71 | 83 | 78 |
| female | group B | some college | standard | completed | 88 | 95 | 92 |
| male | group B | some college | free/reduced | none | 40 | 43 | 39 |
| male | group D | high school | free/reduced | completed | 64 | 64 | 67 |
| female | group B | high school | free/reduced | none | 38 | 60 | 50 |
| male | group C | associate's degree | standard | none | 58 | 54 | 52 |
| male | group D | associate's degree | standard | none | 40 | 52 | 43 |
| female | group B | high school | standard | none | 65 | 81 | 73 |
| male | group A | some college | standard | completed | 78 | 72 | 70 |
| female | group A | master's degree | standard | none | 50 | 53 | 58 |
| female | group C | some high school | standard | none | 69 | 75 | 78 |
| male | group C | high school | standard | none | 88 | 89 | 86 |
| female | group B | some high school | free/reduced | none | 18 | 32 | 28 |
| male | group C | master's degree | free/reduced | completed | 46 | 42 | 46 |
| female | group C | associate's degree | free/reduced | none | 54 | 58 | 61 |
| male | group D | high school | standard | none | 66 | 69 | 63 |
| female | group B | some college | free/reduced | completed | 65 | 75 | 70 |

The proposed system is evaluated using four metirce accuary, precision, recall and F1 score. The accuracy is defined by the following formula to measure the proportion of correctly classified instances ( true positive and true negative ) out of the total instances.

$$\text{Accuracy}= \frac{TP+TN}{TP+TN+FP+FN} \text{ ... (1)}$$

While TP is True Positivem TN is True Negative, FP is False Positive, and FN is False Negative. The precision is defined in formula (2) to measure the proportion of correctly true positive out of all predicted positive instances. High value indicates a low false positive rate

$$\text{Precision}= \frac{TP}{TP+FP} \text{ ... (2)}$$

Recall is defined in formula (3) to measure the proportion of correctly true positive out of all actual positive instances. High value indicates the model capture the actual positives.

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

F1 score is a balance between the precision and recall, which is defined in the formula (4). F1-score between 0 and 1, where 1 indicate perfect precision and recall.

$$\text{F1-Score}= 2. \frac{Precsion*recall}{Precsion+recall} \text{ ... (3)}$$

When it comes to predicting student performance, The models were evaluated using four metrics: accuracy, precision, recall and F 1 score. These metrics provide insights, into the strengths and weaknesses for each model helping determine, which ones are most effective and it distinguishing between low/ how performing students using 80% training and 20% testing . Logistic Regression achieves high accuracy rate of 99.50%, precision rate of 99.10%, recall rate of 100.00%, and an F1 score of 99.55%. These statistics suggest that Logistic Regression not did well overall but showed great consistency in recognizing both high and low achievers effectively. The precision score of 99.10 % indicates that all students identified as performers were truly high achievers compared with rate of misclassification. Moreover, the perfect recall rate of 100% demonstrates that Logistic Regression effectively recognized all achievers without missing any top performing students. The impressive F1 score of 99.55 % highlights the model's ability to balance, between precision and recall making it a reliable option for environments. Furthermore, its interpretability provides insight by helping educators identify which factors have the significant impact, on predictions. Figure (3) shows the main confusion matrix for logistic regression

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

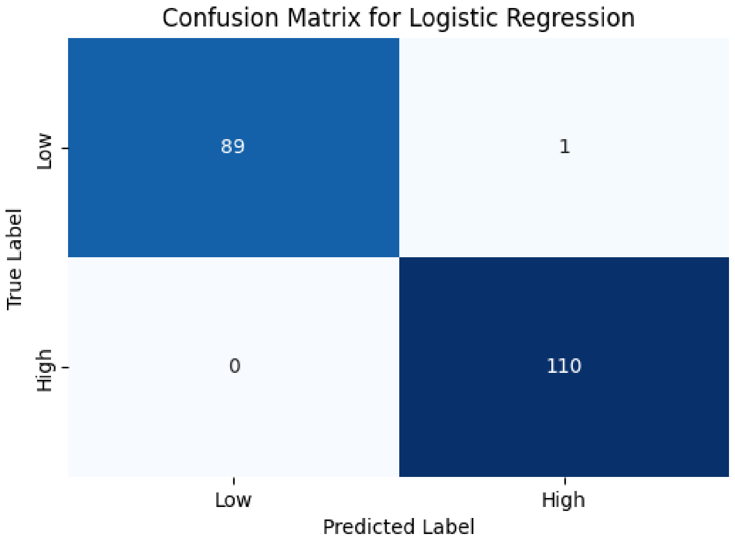

**Figure 3:** Logistic regression confusion matrix

The Decision Tree classifier may not be as exact as Logistic Regression. Still brought results with 97% accuracy and impressive scores across for precision of (96%), recall (98%), and F1 score of (97%). Its interpretability also proves beneficial, for educators looking to understand how certain factors impact student performance. Even though the Decision Trees accuracy is a bit lower, than that of Logistic Regression its precision and recall scores are quite impressive showing its ability to performer students accurately. Yet, the Decision Trees inclination excels in handling nonlinear connections with more effectively. Figure (4) shows the main confusion matrix for decision tree

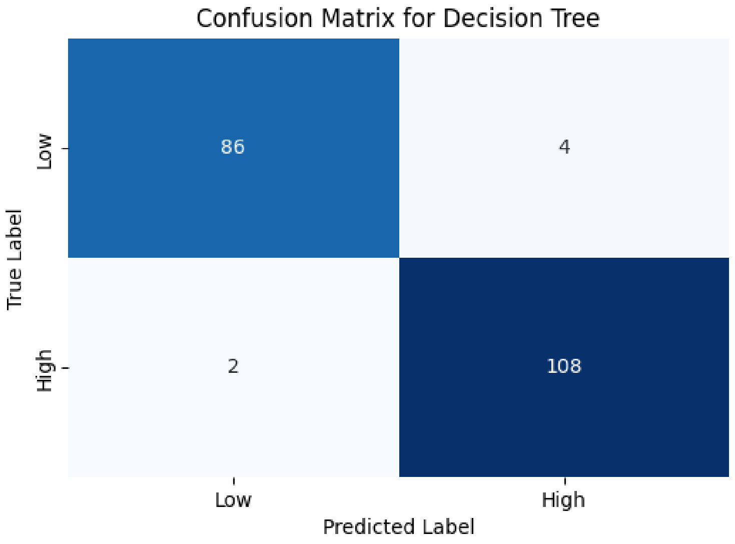

**Figure 4:** Decision Tree confusion matrix

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Random Forest also showed performance, with an accuracy of 98%, precision of 99%, recall of 97%, and F1 score of 98%. Random Forest is created as a model to improve its capacity, in capturing complex data patterns and preventing overfitting issues efficiently. The remarkable precision score (99%) signifies its proficiency in recognizing performers while minimizing the occurrence of false positives and ensuring that resources are directed towards students who are genuinely excelling in their academics. The small decrease in accuracy (97.27%) in comparison to Logistic Regression indicates that even though the results of Random Forest is overall, it may mistakenly categorize some students as low achieves. . Howeverc, its rounded $F_1$ score (98.17%) which comes to understand the impact of particular features. Figure (5) shows the main confusion matrix for random forest.

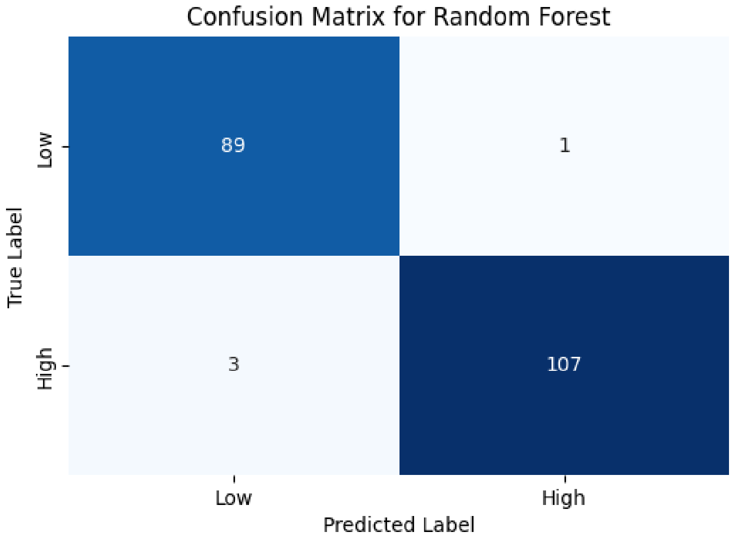

**Figure 5:** Random Forest confusion matrix

The Support Vector Machine (SVM) achieves an accuracy of 99%, a precision of 98%, a recall of 100%, and an F1 score of 99%. The high precision of recall scores of the SVM show its strength in recognizing performers without missing any students in the process for student support initiatives. The models impressive F1 score of 99% indicates its reliability by balancing between accuracy and completeness. SVM models may lack transparency compared with decision trees or logistic regression models which might limit their utility for educators seeking clarity. Figure (6) shows the main confusion matrix for Support Vector Machine.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

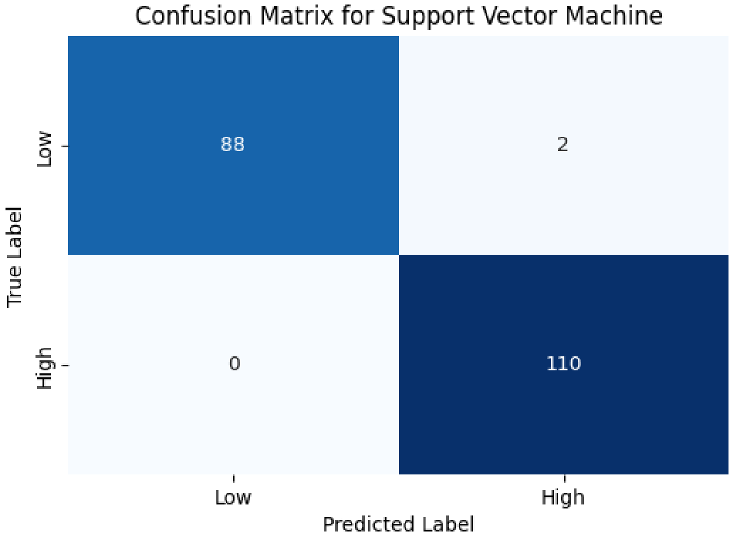

**Figure 6:** Support Vector Machine confusion matrix

The K neighbors (KNN) algorithm also demonstrated results with an accuracy rate of 99%, precision of 99%, recall of 99%, and F1 score of 99%. The balanced metrics of KNN suggest its ability to categorize students accurately with mistakes making it a reliable choice for predicting process. However, the effectiveness of KNN is influenced by the selection of neighbors and data distribution. Its performance could reduce when dealing with datasets. Figure (7) shows the main confusion matrix for KNN.

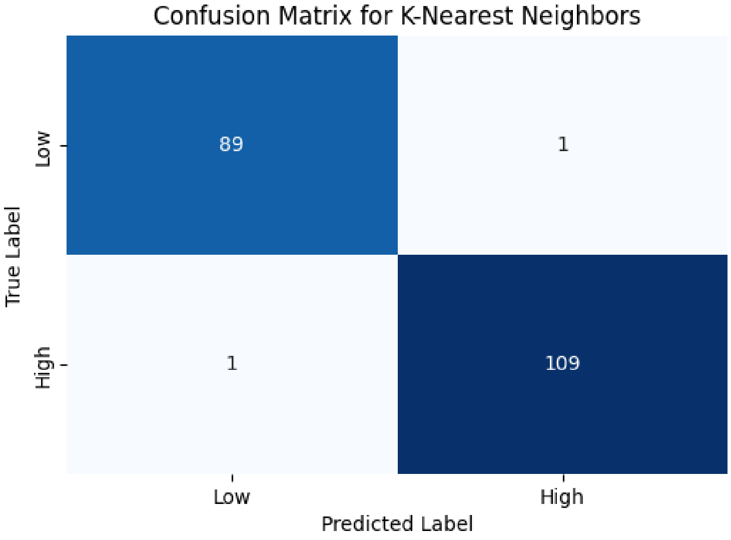

Figure 7: KNN confusion matrix

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مـجلـة جـامعـة بـابـل للعلــوم الصــرفـة والتطــبيقيـة مـجلـة جـامعـة بـابـل للعلــوم الصــرفـة والتطــبيقيـة مـجلـة جـامعـة بـابـل

In conclusion despite performing than the other methods the Naive Bayes classifier still delivered satisfactory outcomes with an accuracy of 96%, a precision and recall of 96% each and an F1 score also, at 96%. Naive Bayes relies on the assumption of feature independence, which might not be entirely applicable in this dataset given the interdependence between academic variables. This particular assumption could be a contributing factor, to its accuracy. Somewhat balanced yet decreased precision and recall rates. Despite this fact Naive Bayes is still a choice that can be quite useful when emphasizing interpretability and speed, over accuracy. Figure (8) shows the main confusion matrix for KNN.

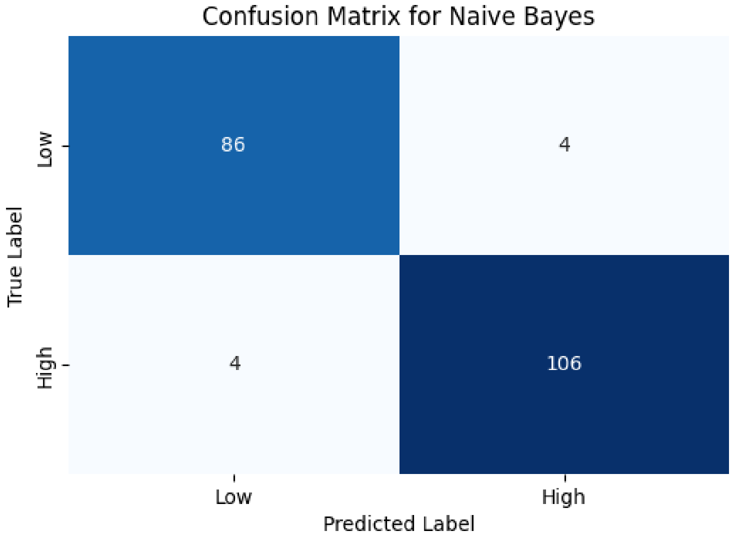

**Figure 8:** Naïve Bays confusion matrix

The summarized chart data in figure (9) for all classifiers was showed that the Logistic Regression, SVM, and K nearest Neighbors have almost perfect scores for accuracy, precision, recall and F1 score with balanced performance. Random Forest comes next with performance across all measures but lower recall. Decision Tree and Naive Bayes show performance in comparison. They are still quite effective especially in situations where interpretability or computational efficiency are important considerations.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

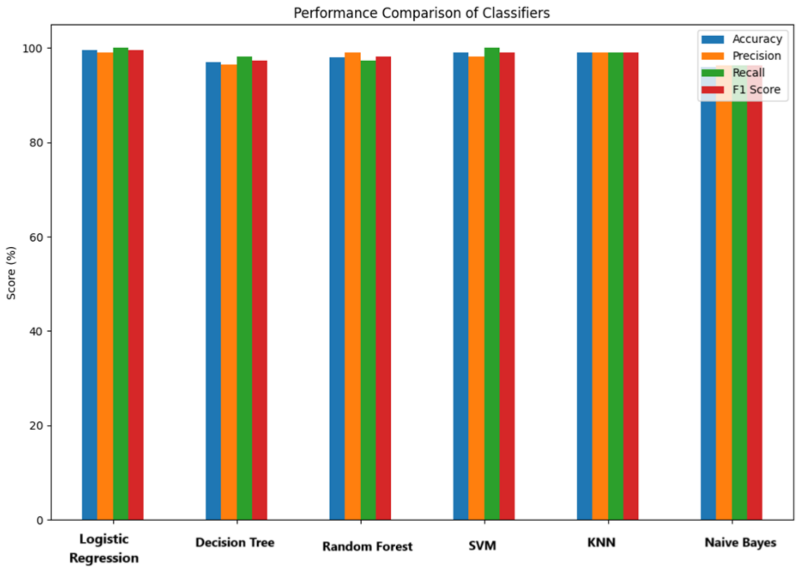

**Figure 9:** Comparison of Student Performance results for different classifiers

Table (2) shows the main summarized results for all machine learning models to predict the student performance.

**Table (2): The results of Machine learning Models**

| Method | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic regression | 99.50% | 99.10% | 100% | 99.55% |
| Decision Tree | 97% | 96% | 98% | 97% |
| Random Forest | 98% | 99% | 97% | 98% |
| Support Vector Machine | 99% | 98% | 100% | 99% |
| K-nearest neighbors | 99% | 99% | 99% | 99% |
| Naïve Bayes | 96% | 96% | 96% | 96% |

## CONCLUSION

Machine learning outperforms excellent results using logistic Regression, SVM, K-nearest neighbors, which standing out as the choice for classification, which outperformed models in terms of accuracy, precision, recall, and F1 score showcasing its dependable predictive capabilities in distinguish between high and low performance. Similarly, Random Forest, decision tree and Naïve bayes yielded results with a lower recall rate but offered valuable insights into feature importance

## Conflict of interests

There are non-conflicts of interest.

## References

[1] S. Alturki, N. Alturki, and H. Stuckenschmidt, "Using Educational Data Mining To Predict Students' Academic Performance For Applying Early Interventions," *Journal of Information Technology Education: Innovations in Practice*, vol. 20, pp. 121–137, 2021, doi: 10.28945/4835.

[2] R. H. Ali, "Educational Data Mining For Predicting Academic Student Performance Using Active Classification," *Iraqi Journal of Science*, vol. 63, no. 9, pp. 3954–3965, 2022, doi: 10.24996/ijs.2022.63.9.27.

[3] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

[4] S. Hussain *et al.*, "Significance of Education Data Mining in Student's Academic Performance Prediction and Analysis," Published|, 2023.

[5] E. Najjar and A. M. Breesam, "Supervised Machine Learning a Brief Survey of Approaches," *Al-Iraqia Journal of Scientific Engineering Research*, vol. 2, no. 4, Jan. 2024, doi: 10.58564/ijser.2.4.2023.121.

[6] Y. Fu *et al.*, "Using machine learning algorithms based on patient admission laboratory parameters to predict adverse outcomes in COVID-19 patients," *Heliyon*, vol. 10, no. 9, May 2024, doi: 10.1016/j.heliyon.2024.e29981.

[7] A. Balboa, A. Cuesta, J. González-Villa, G. Ortiz, and D. Alvear, "Logistic regression vs machine learning to predict evacuation decisions in fire alarm situations," *Saf Sci*, vol. 174, Jun. 2024, doi: 10.1016/j.ssci.2024.106485.

[8] R. Pugliese, S. Regondi, and R. Marini, "Machine learning-based approach: Global trends, research directions, and regulatory standpoints," Dec. 01, 2021, *KeAi Communications Co.* doi: 10.1016/j.dsm.2021.12.002.

[9] S. A. Alwarthan, N. Aslam, and I. U. Khan, "Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review," 2022, *Hindawi Limited*. doi: 10.1155/2022/8924028.

[10] O. A. Montesinos López, A. Montesinos López, and J. Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing, 2022. doi: 10.1007/978-3-030-89010-0.

[11] C. Gürol, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal, "Binary Classification Performance Measures/ Metrics: A Comprehensive Visualized Roadmap to Gain New Insights," in *2nd International Conference on Computer Science and Engineering : Antalya-Türkiye 5-8 Ekim (October) 2017*, IEEE, 2017.

[12] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim, and P. Garcia, "Using educational data mining techniques to increase the prediction accuracy of student academic performance," *Information and Learning Science*, vol. 120, no. 7–8, pp. 451–467, Sep. 2019, doi: 10.1108/ILS-03-2019-0017.

[13] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," Jan. 01, 2021, *MDPI AG*. doi: 10.3390/app11010237.

[14]E. Okewu, P. Adewole, S. Misra, R. Maskeliunas, and R. Damasevicius, “Artificial Neural Networks for Educational Data Mining in Higher Education: A Systematic Literature Review,” *Applied Artificial Intelligence*, vol. 35, no. 13, pp. 983–1021, 2021, doi: 10.1080/08839514.2021.1922847.

[15]M. Arashpour *et al.*, “Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization,” *Computer Applications in Engineering Education*, vol. 31, no. 1, pp. 83–99, Jan. 2023, doi: 10.1002/cae.22572.

[16]W. Xiao, P. Ji, and J. Hu, “A survey on educational data mining methods used for predicting students’ performance,” May 01, 2022, *John Wiley and Sons Inc*. doi: 10.1002/eng2.12482.

[17]Balqis Albreiki, “Identification of Students At Risk of low Performance by combining rule-Based Models, Enhanced machine Learning, and Knowledge Graph Techniques,” United Arab Emirates University, 2023. [Online]. Available: https://scholarworks.uaeu.ac.ae/all_dissertations

[18]D. Hao, Y. Xiaoqi, and Q. Taoyu, “Hybrid Machine Learning Models Based on CATBoost Classifier for Assessing Students’ Academic Performance,” 2024. [Online]. Available: www.ijacsa.thesai.org

[19]Edmund F. Agyemang *et al.*, “Predicting Students’ Academic Performance Via Machine Learning Algorithms: An Empirical Review and Practical Application,” *Computer Engineering and Intelligent Systems*, Sep. 2024, doi: 10.7176/CEIS/15-1-09.

[20]M. D. Adane, J. K. Deku, and E. K. Asare, “Performance Analysis of Machine Learning Algorithms in Prediction of Student Academic Performance,” *Journal of Advances in Mathematics and Computer Science*, vol. 38, no. 5, pp. 74–86, Mar. 2023, doi: 10.9734/jamcs/2023/v38i51762.

[21]N. Bakar, N. A. Mohamad Rejeni, and A. Nyuak, “Machine Learning for Predicting Students’ Academic Achievement Based on Learning Style and Academic Results,” *International Journal of Innovation and Industrial Revolution*, vol. 5, no. 15, pp. 120–130, Dec. 2023, doi: 10.35631/ijirev.515013.

[22] J. H. Guanin-Fajardo, J. Guaña-Moya, and J. Casillas, “Predicting Academic Success of College Students Using Machine Learning Techniques,” *Data (Basel)*, vol. 9, no. 4, Apr. 2024, doi: 10.3390/data9040060.

[23] Seshapanpu, Jakki ; "Students Performance in Exams," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams. [Accessed: 4-Oct-2024].

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل

## الخلاصة

المقدمة:

تستفيد الأنظمة التعليمية بشكل متزايد من الأساليب التحليلية لتحسين الجودة الأكاديمية للطلاب. يعد التنبؤ بأداء الطلاب مجالًا مهمًا باستخدام استخراج البيانات والتعلم الآلي يمكن أن يوفر رؤى مهمة. تؤكد هذه الدراسة على إمكانات التعلم الآلي في البيئات من خلال توفير المعلومات للتدخلات القائمة على البيانات والتي تهدف إلى تحسين نتائج الطلاب ودعم الاستراتيجيات التعليمية.

طرائق العمل:

تضمن مرحلة المعالجة الأولية هذه ملاءمة النموذج. المرحلة الثانية للأمام لخطوة التدريب باستخدام المصنفات، مثل شجرة القرار الانحدار اللوجستي وآلة دعم المتجه (SVM) و Kأقرب الجيران (KNN) و Naive Bayes.

النتائج:

تم الحصول على نتائج الطلاب واختبارها باستخدام ستة نماذج تصنيف؛ جاء الانحدار اللوجستي في المقدمة بدقة 99٪؛ يليه SVM و K أقرب الجيران بنسبة 99٪؛ الغابة العشوائية التي تم إجراؤها بنسبة 98٪؛ شجرة القرار بنسبة 97٪؛ و Naive Bayesبنسبة 96٪. أظهرت نتائج درجات الدقة والتذكر أن النماذج حققت أداءً جيدًا، حيث حقق الانحدار اللوجستي وأقرب الجيران k معدلات مثالية بنسبة 100% تقريبًا. تكشف نتائج خرائط الحرارة لمصفوفة الارتباك لمقارنة الأداء عن فعالية المصنفات القائمة على المجموعة والهامش في هذه المهمة.

الاستنتاجات:

تسلط النتائج الضوء على كيفية مساعدة التعلم الآلي للمدارس في اتخاذ القرارات بشأن دعم الطلاب وتوقع احتياجات الطلاب بشكل فعال. بالنسبة للعمل المستقبلي، سيأخذ العمل المقترح في الاعتبار مقاييس مثل بيانات السلوك والمشاركة وتحديث النماذج باستمرار لتحسين الدقة لتحسين أداء الطلاب.

الكلمات المفتاحية: التعدين في البيانات التعليمية(EDM) ، شجرة القرار، SVM، KNNو Naive Bayes.