*Nanotechnology Perceptions*
*ISSN 1660-6795*
www.nano-ntp.com

# Developing A Machine Learning Predictive Model For Learning Analytics In Higher Education

**Hussam Mohammed Alamoudi[*] [1 2], Adel Bahaddad [1]**

[1] *Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
[2] *Department of Management Information System, College of Business Administration, Jazan University, Jazan, Saudi Arabia*
** Corresponding Author: halamoudi@jazanu.edu.sa /*

This study presents the development of a Predictive Learning Analytics Model (PLAM) utilizing Educational Data Mining (EDM) and machine learning algorithms to enhance Teaching and Learning Outcomes (TLOs) at Jazan University's College of Business Administration (JCBA). The research aimed to detect hotspots and predict academic failure among JCBA students to implement proactive interventions. Data from surveys and e-registers involving demographic and academic information from 212 participants over three academic years (2020-2023) were analyzed. An .arff file was created from the collected data, featuring 19 attributes plus one class, and analyzed using the WEKA tool with various classification algorithms. The main findings indicate that the Random Forest classifier achieved the highest accuracy, significantly predicting student performance and identifying at-risk students. This model provides valuable insights for data-driven educational strategies, demonstrating the effectiveness of machine learning in improving academic outcomes. The choice of Jazan University was motivated by the need to address educational challenges and leverage data analytics to enhance student success.

**Keywords:** Data Analysis, Student Engagement, Algorithm Evaluation, Academic Performance, Educational Insights, Predictive Analytics.

## I. Introduction

In an era where data is ubiquitously collected and analyzed across various domains, education stands out as a field ripe for transformative insights and improvements through data analytics. The vast amounts of data generated within educational settings hold the key to unlocking potential advancements in teaching methodologies, learning outcomes, and overall academic performance. Educational data, encompassing grades, attendance records, behavioral logs, and more, offers a rich tapestry of information that, when properly analyzed, can lead to significant enhancements in how educational institutions function and how students learn. This work introduces a Predictive Learning Analytics Model (PLAM) utilizing the Educational Data Mining (EDM) process aimed at enhancing Teaching and Learning Outcomes (TLOs) at Jazan University's College of Business Administration (CBA). The urgency for such an exploration

stem from the ongoing challenge of optimizing educational practices to meet the ever-evolving demands of the global workforce and the individual needs of students. With education becoming increasingly competitive and the job market demanding higher levels of skill and adaptability, institutions must leverage all available resources, including data, to stay ahead. The backdrop of this study is set against the burgeoning field of EDM, where the collection, analysis, and application of educational data can reveal patterns, predict outcomes, and inform strategic educational decisions. EDM integrates principles from data mining, machine learning, and educational theory to extract meaningful insights from educational data. This multidisciplinary approach not only aids in understanding current educational dynamics but also in forecasting future trends and potential issues.

The emergence of PLAMs represents a sophisticated approach to harnessing the power of machine learning algorithms and analytics to forecast and improve educational trajectories. Machine learning, with its ability to handle large datasets and uncover complex relationships, is particularly well-suited for this task. PLAMs can analyze a multitude of variables simultaneously, identifying which factors most significantly impact student success and predicting which students are at risk of poor performance. This work is motivated by the potential of such models to revolutionize the educational landscape by providing actionable insights that can personalize learning, preemptively address student needs, and ultimately elevate the quality of education delivered. Personalized learning, informed by data-driven insights, can adapt educational content to the individual needs of students, thereby enhancing engagement and improving outcomes. Early identification of students at risk allows for timely interventions, which can mitigate potential academic failures and support students in achieving their full potential.

This study is anchored in the context of Jazan University's CBA, serving as a case study for the application and evaluation of the proposed PLAM. By focusing on a specific academic setting, this study aims to offer a detailed analysis of the PLAM's efficacy, the nuances of its implementation, and the broader implications for educational practice. Jazan University was chosen due to its strategic importance in the region and its commitment to improving educational standards through innovative approaches. This context provides a fertile ground for testing and validating the PLAM, with potential lessons that can be extended to other educational institutions. The research questions guiding this study delve into the development of the PLAM, its effectiveness in improving TLOs, and its potential impact on educational strategies and policies. Specifically, the study seeks to answer how a PLAM can be developed using EDM and machine learning algorithms, how effective it is in predicting student performance, and what implications it has for educational practice. By addressing these questions, the study aims to contribute to the growing body of knowledge in EDM and provide practical insights that can help educators and policymakers improve educational outcomes. To comprehensively address these objectives, the structure of this paper is as follows: Section (II) **The Comprehensive Theoretical Basis** provides a theoretical foundation and context by examining previous studies and current knowledge in the field; Section (III) **Methodology** details the research design, data collection, and analysis procedures; Section (IV) **Results and Analysis** presents the outcomes of the PLAM implementation and finally Section (V) **Discussion and Recommendations** interprets the results and offers practical suggestions for

educational practice. The Conclusion summarizes the findings and their implications and suggests directions for future research. This research represents a critical step towards leveraging advanced data analytics in education. It underscores the importance of data-driven decision-making in modern educational environments and highlights the transformative potential of machine learning in understanding and enhancing student learning experiences.

## II. The Comprehensive Theoretical Basis

### 1) Educational Data Mining Process

Educational Data Mining (EDM) applies data mining techniques to educational data, involving key steps such as data collection, preprocessing, exploratory data analysis (EDA), data modeling, and interpretation of results [1]. Data collection gathers information from learning management systems, assessments, and student information systems. Moreover, preprocessing ensures data quality by cleaning and transforming it. EDA uses visualization and descriptive statistics to understand data characteristics and relationships, guiding hypotheses for data modeling [2]. However, data modeling employs statistical and machine learning techniques to predict student performance and uncover patterns. Techniques include regression analysis, clustering, classification, and association rule mining, chosen based on research questions and data characteristics. Furthermore, the interpretation of results involves analyzing and validating findings to ensure they support the initial hypotheses [3-4]. Since EDM aids evidence-based decision-making, it helps identify at-risk students and evaluate educational programs' effectiveness [5-7]. Additionally, it supports adaptive learning systems, dynamically tailoring instructional materials based on learner data, thereby enhancing personalized learning experiences [8].

### 2) Predictive Learning Analytics Model

Predictive Learning Analytics Models (PLAMs) use data mining and machine learning to predict student performance. Key components include data collection, preprocessing, feature engineering, model training, evaluation, and result interpretation [9]. Since data collection involves gathering information from educational sources, preprocessing cleans and integrates the data. Moreover, feature engineering selects relevant variables for the model [10]. Model training applies algorithms like decision trees, logistic regression, support vector machines, and neural networks to learn relationships between input features and target variables. However, model evaluation assesses performance using metrics such as accuracy and F1 score to ensure reliability and generalizability. Interpretation of results generates insights for instructional design, personalized learning, and identifying at-risk students [11]. PLAMs enhance student outcomes by providing timely predictions and enabling early interventions. Additionally, they support personalized learning by tailoring instruction to individual needs and informing policy and resource allocation decisions. By analyzing educational data, PLAMs can improve student engagement, retention, and performance [12].

### 3) Teaching and Learning Outcomes (TLOs)

Teaching and Learning Outcomes (TLOs) are essential in shaping educational experiences, guiding instructional practices to align with institutional goals and student needs. Jazan University's College of Business Administration (CBA) has implemented TLOs to enhance teaching and learning processes. Since TLOs promote critical thinking, problem-solving, communication, and teamwork skills, strategies include case studies, group discussions, and technology-enhanced learning. International benchmarks demonstrate EDM's impact on TLOs improvement. For instance, Carnegie Mellon's Open Learning Initiative (OLI) [13], MIT's Data-driven Feedback for Learning (D4L) [14], Stanford's Educational Program for Gifted Youth (EPGY), UC Berkeley's Online Education Initiative (OEI) [15], and Harvard's HarvardX use EDM [16] to personalize learning, provide feedback, and improve engagement and performance. Hussain et al. (2018) [17] investigated EDM techniques on medical college admissions data, finding neural networks most effective in classification tasks. Their study highlights the importance of multi-year datasets for generalizability, a focus of this research at Jazan University.

### 4) Relationship between EDM and TLOs Improvement

Since EDM uses data analysis to inform decision-making, it enhances educational practices and TLOs [18]. By analyzing student performance and learning analytics data, educators can tailor instructional strategies to meet individual and group needs, thereby improving TLOs [19]. Moreover, predictive analytics identifies at-risk students early, enabling timely interventions and support [20]. These benchmark models demonstrate EDM's role in improving TLOs by informing instructional practices, personalized interventions, and evidence-based decisions. Consequently, EDM positively impacts student engagement, retention, performance, and overall learning experience, which are crucial for modern educational settings [20-21].

## III. Methodology

The methodology of this research is designed to systematically develop a Predictive Learning Analytics Model (PLAM) using Educational Data Mining (EDM) techniques at Jazan University. The foundational step involves the collection of data from two primary sources:

- **The electronic register** (e-register) which contain students' enrollment details, attendance, grades, course registrations, and all academic activities…
- **Tailored surveys** providing demographic and academic information.

Both of which have been collected with a focus on the College of Business Administration. The collected data is poised to offer a comprehensive overview of the students' academic engagements, capturing a wide array of variables, from demographic details to academic performance indicators. Once the data is amassed, it is converted into an Attribute-Relation File Format (.arff); a format compatible with the Waikato Environment for Knowledge Analysis (WEKA), a suite of machine learning software. This conversion is critical as it structures the data into a format that is readily analyzable by the data mining tools. The pre-processed data then undergoes a rigorous feature selection process within WEKA, aiming to

identify the most predictive attributes that contribute to academic performance. Feature selection is a pivotal phase where irrelevant or redundant data is discarded, enhancing the efficiency and accuracy of the predictive model. Post feature selection, the data is divided into two distinct sets:

- The training dataset: which is used to build the PLAM,
- The test dataset: which serves to evaluate its predictive prowess.

The research then progresses into the model development phase, employing various classification algorithms to train the PLAM. The algorithms chosen uses a spectrum of statistical and machine learning techniques, including but not limited to, Bayes such as Naïve Bayes, various functions such as Linear Regression, Simple Linear, Gaussian, rule-based algorithms such as ZeroR, OneR, Decision Table, M5, and decision trees such as J48, Random Forest. This diverse array of algorithms is methodically applied to the training dataset to establish the most effective technique for the predictive model based on the patterns it uncovers. Once the model is developed, it is imperative to assess its performance. This evaluation is conducted through the application of the model to the test dataset, with the outcomes being scrutinized against a set of performance metrics, such as accuracy, precision, recall, and the F-measure. These metrics provide a quantitative assessment of the model's ability to accurately predict student performance. Furthermore, the results are visualized using various techniques to facilitate intuitive understanding and to communicate the findings to stakeholders who might not have a technical background. Graphical representations such as confusion matrices, ROC curves, and performance graphs are employed to illustrate the model's predictive accuracy and the importance of the selected features such as accuracy, precision, recall.

The culmination of this research methodology will not only yield a robust PLAM but also provide insightful patterns and relationships within the educational data. These insights can significantly impact how educational support is administered, paving the way for targeted interventions and informed decision-making that foster academic success at Jazan University. Through careful adherence to this structured methodology, this research aims to make a valuable contribution to the field of Educational Data Mining and to the academic community at Jazan University. The following Figure 1 summarizes the experimental framework and the different working steps.

**1) Data collection**

Three different data types are collected by means of surveys and e-register:

- Demographic Information (DI)
- Academic Information (AI)
- Academic Performance (Exams score)

The survey was conducted for 15 days starting from June 12th till June 26th, 2023, and contains 19 questions and divided into two main sections:

- Demographic Information (DI): providing 13 parameters called attributes.
- Academic Information (AI): providing 6 parameters also called attributes.

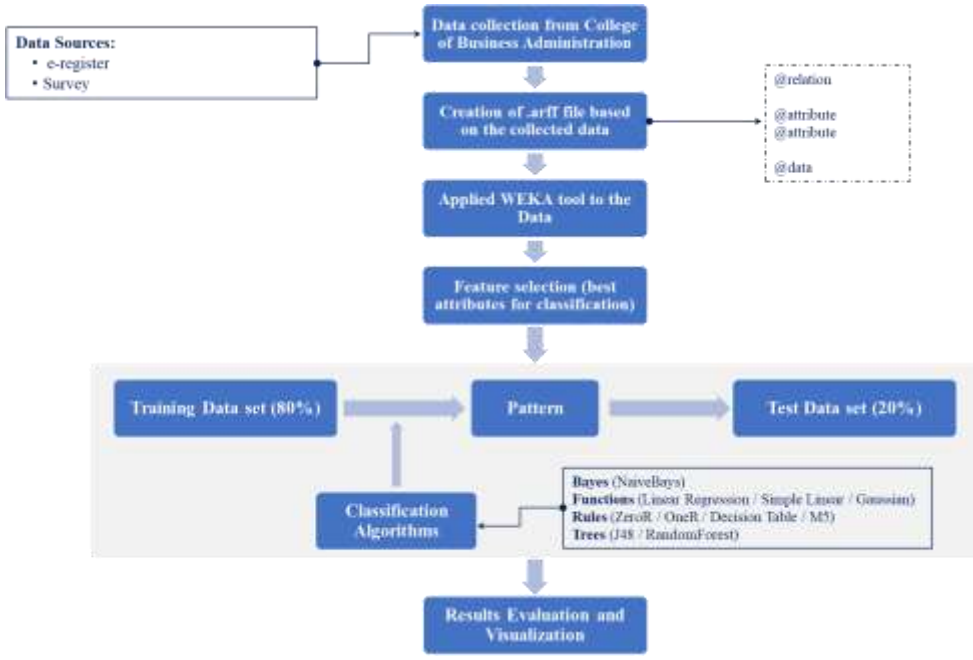


**Figure 1 – The experimental framework**

The following Table 1 summarizes these parameters / attributes along with their abbreviations used in the model development process.

**Table 1 – List of used demographic and academic attributes**

| # | Attribute Name | Description | Attribute Value |
|---|---|---|---|
| 1 | G | Gender | Male = "M" / Female = "F" |
| 2 | MS | Marital Status | Single = "S / Married = "M" / Divorced = "D" |
| 3 | AD | Place of residence | Urban = "I" / Rural = "0" |
| 4 | FES | Family Economic Status | Low = "L" / Meduim = "M" / High = "H" |
| 5 | FS | Family size | Small = "S" / Meduim = "M" / Large = "L" |
| 6 | FLE | Father Level of Education | {0, 1, 2, 3, 4, 5, 6, 7} – for more details see survey |
| 7 | MLE | Mather Level of Education | {0, 1, 2, 3, 4, 5, 6, 7} – for more details see survey |

| # | Attribute Name | Description | Attribute Value |
|---|---|---|---|
| 8 | FO | Father Occupation | {0, 1, 2, 3, 4, 5, 6} – for more details see survey |
| 9 | MO | Mather Occupation | {0, 1, 2, 3, 4, 5, 6} – for more details see survey |
| 10 | TH | Type of Housing | At the University Campus = "U" / Private = "P" |
| 11 | TT | Travel time between college and home | Very Short = "VS" / Short = "S" / Meduim = "M" / Long = "L" |
| 12 | NF | Number size of friends | Small = "S" / Meduim = "M" / Large = "L" |
| 13 | MJ | Major | {MGIS, ADMN, FIBA, ACCT, MRKT} |
| 14 | YS | Year of Study | 1ST year = "1" / 2nd year = "2" / 3rd year = "3" / Final year = "4" |
| 15 | ECV | Extracurricular activities or clubs | YES = Y / NO =N |
| 16 | HFS | Hours per week for studying | Very Short = "VS" / Short = "S" / Meduim = "M" / Long = "L" / Very Long = "VL" |
| 17 | AST | Academic support or tutoring (Frequency) | {0, 1, 2, 3, 4} – for more details see survey |
| 18 | LRU | College/university library resources Utilization | {0, 1, 2, 3, 4} – for more details see survey |
| 19 | RFS | Educational resources and facilities satisfaction | {0, 1, 2, 3} – for more details see survey |

The e-register will provide us with the academic performance related to JCBA student's results (Pass or Fail). This will be the output of the model, also called "Class".

**2) Data Analysis**

The analysis of the survey data collected over the three last academic years (2020/2021, 2021/2022, and 2022/2023) from Jazan University's JCBA provides a detailed insight into the academic environment and the demographics of the participants. Out of the 212 participants surveyed, a significant majority of 70.3% were male, while 29.7% were female. This gender distribution reflects a higher male engagement in the survey, which could be indicative of the gender ratio at JCBA or possibly a higher response rate among male students to the survey. The participants were distributed across academic years, with 25% in their first year of study. This indicates a good engagement level among new students, which is crucial as this group is adapting to university life and their experiences can provide fresh insights into the introductory academic environment and support services. Second-year students constituted 22.1%, suggesting a slight drop in participation or population as students' progress in their studies.

The third-year students, who represented the largest group at 34%, might provide the most substantive feedback regarding the curriculum as they are deeply involved in their major-specific courses by this point. Finally, 18.9% of participants were in their final year of study, a critical stage where students are preparing to transition to the workforce or further studies. The dataset was prudently sorted, with 159 selected answers deemed suitable for analysis. Among these 159 selected entries 119 were "Pass" labeled and the remaining 40 were "Fail" labeled. Participants from various departments within the JCBA were surveyed, including 19 students from Finance and Banking department (FIBA), 20 students from marketing departments (MRKT), 28 students from accounting department (ACCT), 46 students from management information systems department (MGIS), and 46 students from business administration department (ADMN). This diverse representation from different departments is critical as it ensures that the model accounts for a wide range of academic experiences and challenges specific to each discipline. Of these, a substantial 80% (128 responses) were allocated for training the predictive model. This substantial proportion for the training set is aligned with common practices in machine learning, allowing for the development of a robust model capable of generalizing from the learned patterns. The remaining 20% (31 responses) were reserved for testing and validating the model. This split ensures that the model's predictive power is assessed on unseen data, providing a measure of its effectiveness and accuracy.

### 3) Model development and classification algorithms

In the model development phase of this research, we applied a suite of classification algorithms to the training data, with the objective of identifying the model that yields the highest accuracy in predicting the academic success of students at Jazan University's College of Business Administration. The algorithms selected for this task included a variety of approaches (Bayes, Rules, Functions and Trees), each with unique strengths and assumptions about the data.

**Table 2 – Used classification algorithm approaches and types and their advantages and disadvantages**

| Classification Algorithm Approaches | Classification Algorithm Types | Advantage | Disadvantage |
|---|---|---|---|
| **Bayes** | Naive Bayes | • Simple and computationally efficient.<br>• Works well with high-dimensional data<br>• Handles missing data effectively. | • Assumes independence between features, which may not hold true in all cases.<br>• Can be sensitive to feature correlations. |
| **Rules** | OneR | • Easy to interpret and understand.<br>• Suitable for small datasets and quick prototyping. | • May not capture complex relationships in the data. |

| Classification Algorithm Approaches | Classification Algorithm Types | Advantage | Disadvantage |
|---|---|---|---|
| | | • Computationally efficient. | • Relatively simplistic compared to other algorithms. |
| | Decision Table | • Provides an explicit representation of decision rules.<br>• Interpretable and easily understandable.<br>• Handles both categorical and numerical data. | • May suffer from data sparsity or large feature spaces.<br>• Limited ability to capture complex decision boundaries. |
| **Functions** | Sequential Minimal | • Efficient for solving large-scale SVM problems.<br>• Effective for high-dimensional data.<br>• Robust against overfitting. | • Requires tuning of parameters like regularization parameter and kernel function.<br>• May be computationally intensive for very large datasets. |
| | Multilayer Perceptron | • Capable of learning complex nonlinear relationships.<br>• Can handle high-dimensional data.<br>• Suitable for a wide range of applications. | • - Sensitive to feature scaling and initialization.<br>• Prone to overfitting, especially with large neural networks. |
| **Trees** | J48 | • - Produces interpretable decision trees.<br>• Handles both categorical and continuous data.<br>• Robust to noisy data.<br>• Efficient for large datasets. | • - May suffer from overfitting, especially with deep trees.<br>• Can be biased towards attributes with many values or levels. |
| | Random Forest | • - Robust against overfitting and noise.<br>• Handles high-dimensional data effectively.<br>• Provides feature importance ranking.<br>• Parallelizable and scalable. | • - Less interpretable compared to single decision trees.<br>• Can be computationally expensive for very large datasets or many trees. |

| **Classification Algorithm Approaches** | **Classification Algorithm Types** | **Advantage** | **Disadvantage** |
|---|---|---|---|
| | Random Tree | • - Simple and computationally efficient.<br>• Reduces overfitting compared to traditional decision trees.<br>• Suitable for real-time prediction tasks. | • - May sacrifice predictive accuracy compared to more sophisticated algorithms.<br>• Less interpretable compared to decision trees. |

The previous Table 2 shows the different Classification algorithm approaches and types. It also provides a concise overview of the advantages and disadvantages of each of the previous described classification algorithm, aiding in the selection of the most appropriate algorithm based on the specific requirements and characteristics of the dataset and task at hand.

**4) Model Evaluation Metrics**

Upon training the models, we proceeded to evaluate their performance on the testing set, which included the remaining 31 responses. The testing phase is crucial as it offers insights into how each model generalizes to new, unseen data — a key indicator of a model's real-world applicability. During the evaluation process, we employed several performance metrics to assess the effectiveness of each classification algorithm in predicting student performance. These metrics included accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC).

- **Accuracy**: It measures the proportion of correctly classified instances out of all instances in the testing set, providing an overall assessment of model performance. It can be calculated using the following equation (1):

$$\textbf{Accuracy}= \frac{\textbf{Number of correctly classified instance}}{\textbf{Total number of instances}} \qquad \text{(Eq.1)}$$

- **Precision**: It quantifies the proportion of true positive predictions out of all positive predictions made by the model, indicating the model's ability to avoid false positives. It can be calculated using the following equation (2):

$$\textbf{Precision}= \frac{\textbf{True positives}}{\textbf{True positives+False positives}} \qquad \text{(Eq.2)}$$

- **Sensitivity**: also known as recall, measures the proportion of true positive predictions out of all actual positive instances in the testing set, reflecting the model's ability to capture all relevant instances. It can be calculated using the following equation (3):

$$\textbf{Recall}= \frac{\textbf{True positives}}{\textbf{True positives+False negatives}} \qquad \text{(Eq.3)}$$

- **F1 Score**: It combines precision and recall into a single metric, providing a balanced assessment of a model's performance. It can be calculated using Precision and Recall metrics as follow (equation 4):

$$\textbf{F1}=2\times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision+Recall}} \qquad \text{(Eq.4)} \qquad \text{(Eq.4)}$$

- **Area Under the Receiver Operating Characteristic Curve (ROC-AUC) Score**: It evaluates the trade-off between true positive rate and false positive rate across different threshold values, offering insights into the model's discriminative ability. Given a set of true positive rates (TPR) and false positive rates (FPR) at various thresholds, the ROC-AUC can be calculated as follows (equation 5):

$$\textbf{ROC}_{\textbf{AUC}} = \sum_{\textbf{i=1}}^{\textbf{n-1}} \frac{\textbf{1}}{\textbf{2}} (\textbf{FPR}_{\textbf{i+1}} - \textbf{FPR}_{\textbf{i}}) \times (\textbf{TPR}_{\textbf{i}} + \textbf{TPR}_{\textbf{i+1}}) \qquad \text{(Eq.5)}$$

Where:

- $FPR_i$ and $TPR_i$ represent the false positive rate and true positive rate at the $i^{th}$ threshold, respectively.
- n is the total number of thresholds.

By systematically evaluating the models using these metrics, one gained a comprehensive understanding of their strengths and weaknesses, enabling informed decisions regarding their suitability for practical deployment in educational settings.

## IV. Results and analysis

We rigorously evaluate the PLAM using a range of classification algorithms, including Naive Bayes, OneR, DecisionTable, Sequential Minimal Optimization (SMO), Multilayer Perceptron, J48, Random Forest, and Random Tree. Each algorithm is assessed based on its ability to predict student performance and inform instructional strategies, providing valuable insights into the effectiveness of the PLAM in enhancing Teaching and Learning Outcomes (TLOs) at CBA.

### 1) Model Accuracy

The accuracy of each model was calculated using equation (Eq.1). Given the results presented in Table 3, one can clearly see that Naive Bayes achieves an accuracy rate of 78%, demonstrating its effectiveness in classifying data despite its simplicity and assumption of feature independence. OneR, with an accuracy rate of 75%, offers a straightforward and interpretable model, though its predictive power may be limited compared to more complex algorithms. DecisionTable performs slightly better, with an accuracy rate of 84%, showcasing its ability to capture intricate decision rules and feature interactions. Notably, SMO, MultilayerPerceptron, RandomForest, and RandomTree all achieve perfect accuracy rates of 100%, indicating their exceptional predictive performance and robustness across various datasets. These algorithms, particularly SMO and MultilayerPerceptron, leverage sophisticated techniques such as neural networks and support vector machines to learn complex patterns and achieve high accuracy. J48 achieves an accuracy rate of 81%, offering a balance between interpretability and performance.

Overall, the accuracy rates provide valuable insights into the strengths and limitations of each classification algorithm, guiding the selection of the most suitable approach based on specific modeling requirements and dataset characteristics.

**Table 3 – Summary of the model evaluation metrics (Accuracy, Precision, Sensitivity/Recal, F1 Score, ROC-AUC Scor) for 8 different evaluated classification algorithms.**

| Classification algorithms type | Accuracy Rate (%) | Precision Rate (%) | Sensitivity/Recall Rate (%) | F1 Score | ROC-AUC Score |
|---|---|---|---|---|---|
| Naïve Bayes | 78% | 82% | 78% | 0.796 | 0.833 |
| OneR | 75% | 75% | 75% | 0.75 | 0.59 |
| DecisionTable | 84% | 86% | 84% | 0.795 | 0.647 |
| SMO (Sequential Minimal Optimization) | 96% | 97% | 96% | 0.968 | 0.917 |
| MultilayerPerceptron | 100% | 100% | 100% | 1 | 1 |
| J48 (an implementation of the C4.5 algorithm) | 81% | 81% | 81% | 0.897 | 0.5 |
| RandomForest | 100% | 100% | 100% | 1 | 1 |
| RandomTree | 100% | 100% | 100% | 1 | 1 |

**2) Model Precision**

The precision of each model was calculated using equation (Eq.2). Given the results presented in Table 3, Naïve Bayes demonstrates a precision rate of 82%, indicating its effectiveness in minimizing false positives despite its simplicity and assumption of feature independence. OneR achieves a precision rate of 75%, which matches its accuracy rate, suggesting consistent performance in correctly identifying positive instances. DecisionTable performs slightly better, with a precision rate of 86%, reflecting its ability to generate accurate predictions while minimizing false positives. Notably, SMO, MultilayerPerceptron, RandomForest, and RandomTree all achieve perfect precision rates of 100%, showcasing their capability to accurately classify positive instances without any false positives. These algorithms leverage sophisticated techniques such as neural networks and ensemble methods to achieve high precision and minimize classification errors. J48 achieves a precision rate of 81%, indicating reliable performance in correctly identifying positive instances. Overall, the precision rates complement the accuracy rates by providing additional insights into the algorithms' ability to minimize false positives, thereby guiding the selection of the most suitable approach based on specific modeling requirements and the importance of precision in the given task.

**3) Model Sensitivity/Recall**

The sensitivity/recall of each model was calculated using equation (Eq.3). Given the results presented in Table 3, NaiveBayes achieves a recall rate of 78%, indicating its effectiveness in capturing a high proportion of positive instances despite its simplicity and assumption of feature independence. OneR matches its recall rate with its accuracy and precision rates, achieving a recall rate of 75%, suggesting consistent performance in correctly identifying positive instances. DecisionTable performs slightly better, with a recall rate of 84%, reflecting its ability to effectively capture actual positive instances while minimizing false negatives. Notably, SMO, MultilayerPerceptron, RandomForest, and RandomTree all achieve perfect recall rates of 100%, showcasing their capability to accurately identify all actual positive instances without any false negatives. These algorithms leverage sophisticated techniques such as neural networks and ensemble methods to achieve high recall and effectively capture positive instances. J48 achieves a recall rate of 81%, indicating reliable performance in capturing actual positive instances. Overall, the recall rates complement the accuracy and precision rates by providing additional insights into the algorithms' ability to effectively identify positive instances, thereby guiding the selection of the most suitable approach based on specific modeling requirements and the importance of recall in the given task.

**4) F1 Score**

The F1 Score of each model was calculated using equation (Eq.4). Given the results presented in Table 3, the F1 score of various classification algorithms provides a balanced measure of their accuracy and robustness, considering both precision and recall. NaiveBayes achieves an F1 score of 0.796, indicating a good balance between precision and recall despite its simplicity. OneR and DecisionTable both exhibit F1 scores of 0.75 and 0.795, respectively, showcasing moderate performance in capturing both precision and recall. Notably, SMO achieves a high F1 score of 0.968, indicating strong performance in achieving both high precision and recall simultaneously. MultilayerPerceptron, RandomForest, and RandomTree all achieve perfect F1

scores of 1, demonstrating exceptional balance between precision and recall and robustness in classifying positive instances. J48, achieves an F1 score of 0.897, indicating reliable performance in capturing both precision and recall effectively. Overall, the F1 scores provide valuable insights into the algorithms' ability to achieve a balance between precision and recall, guiding the selection of the most suitable approach based on specific modeling requirements and the importance of balanced performance in the given task.

**5) Area Under the Receiver Operating Characteristic Curve (ROC-AUC) Score**

The ROC-AUC Score of each model was calculated using equation (Eq.5). Given the results presented in Table 3, NaiveBayes achieves a respectable ROC-AUC score of 0.833, indicating good discriminative ability despite its simplicity. However, OneR and DecisionTable exhibit lower ROC-AUC scores of 0.59 and 0.647, respectively, suggesting suboptimal performance in distinguishing between positive and negative classes. SMO demonstrates a strong ROC-AUC score of 0.917, indicating robust discrimination capabilities. MultilayerPerceptron, RandomForest, and RandomTree achieve perfect ROC-AUC scores of 1, showcasing exceptional discriminative ability and effectiveness in classifying instances. J48, although widely used, exhibits a lower ROC-AUC score of 0.5, indicating poor discrimination between positive and negative classes. Overall, the ROC-AUC scores provide valuable insights into the algorithms' ability to discriminate between classes and guide the selection of the most suitable approach based on specific modeling requirements and the importance of discriminative ability in the given task.

The following Figure 2 summarizes the classification algorithm’s performance metrics; accuracy, precision, recall rates and F1, ROC-AUC scores.

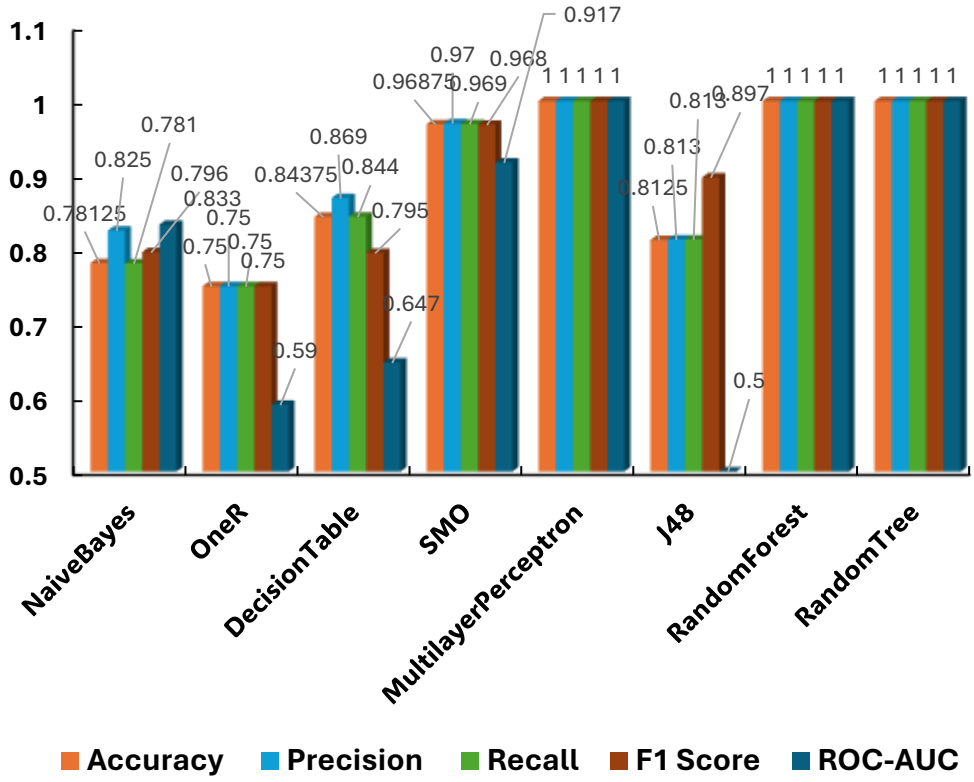


**Figure 2 – Classification algorithm's performance metrics evaluation**

Based on these performance metrics, the RandomTree classification algorithm achieves perfect scores in all metrics, indicating exceptional performance in predicting student outcomes. It exhibits 100% accuracy, precision, recall, F1-score, and ROC-AUC, suggesting robustness and reliability in classifying student performance. While other algorithms may perform well in certain metrics, none achieve the consistent and flawless performance demonstrated by the RandomTree. Therefore, considering its superior performance across all evaluation criteria, the RandomTree can be considered the best classification algorithm for this task.

## V. Discussion and recommendations

In this work, the predictive capabilities of the PLAM were analyzed and potential limitations inherent in the data and modeling approaches were identified. By examining the discrepancies between predicted and actual outcomes, we gain a nuanced understanding of the PLAM's strengths and weaknesses, paving the way for iterative improvements and refinements.

The prediction results show a perfect alignment between actual and predicted outcomes for all performed 32 instances, with 'P' indicating Pass and 'F' indicating Fail. In every case, the predicted outcome matched the actual outcome, as reflected by the consistent

'1' in the error prediction, which signifies correct predictions across the board. This flawless accuracy suggests that the chosen model (Random Tree algorithm) performed exceptionally well on this dataset, achieving 100% prediction accuracy for student performance outcomes.

This perfect alignment between predicted and actual outcomes would usually be highly commendable, suggesting that the chosen model (the Random Tree algorithm, as discussed earlier) is performing with exceptional accuracy on this dataset. However, in practical applications, such flawless performance might be subject to scrutiny, as it is uncommon for a model to achieve 100% accuracy on real-world data due to noise and other factors unless the dataset is very well-defined with clear decision boundaries, or the model has overfit to the training data. It is also crucial to consider the diversity and size of the dataset when evaluating model performance. A small or non-representative dataset might yield high accuracy but fail to generalize well to the broader population. This phenomenon is known as overfitting, where the model learns the training data too well, including its noise and outliers, and does not perform well on unseen data.

Moreover, continuous monitoring and validation with new data are essential to ensure that the model remains accurate over time. In educational settings, factors influencing student success can evolve, so the model might need to be retrained or adjusted to maintain its predictive power. Further analysis could also delve into the confidence levels of predictions, the balance of the classes in the dataset, and the model's performance across different subgroups within the data.

Figure 3 presents the model tree view generated by WEKA software; it provided showcases an intricate decision tree, reflecting a comprehensive model constructed by the Random Tree algorithm. The depth and breadth of the tree (size of the tree 188) are indicative of a rich learning process, where the algorithm has delved into the subtleties of the dataset to carve out a detailed series of decision paths. Each node in this tree serves as a checkpoint that evaluates certain attributes, guiding the way down to a leaf that symbolizes a clear decision.

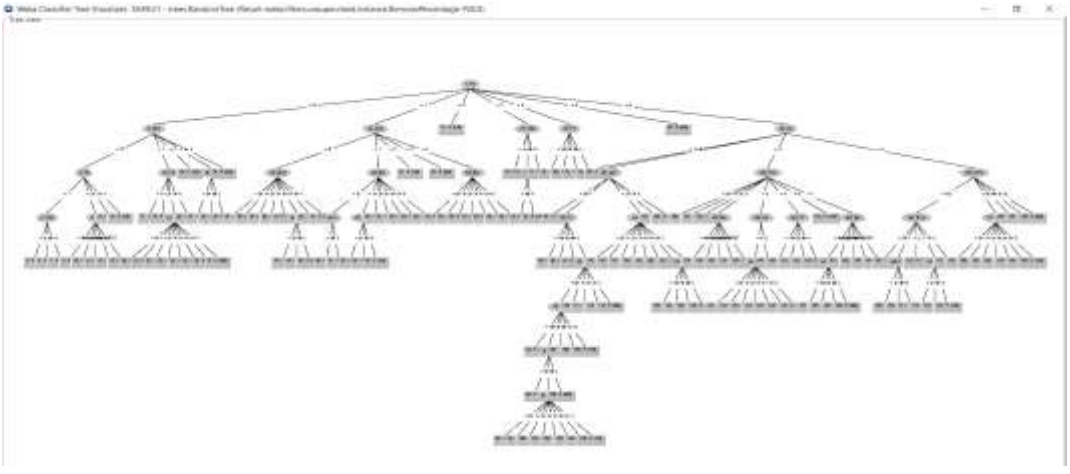

**Figure 3 – Model tree view generated by WEKA software.**

This complexity is a testament to the algorithm's capability to understand and categorize various scenarios, potentially capturing the nuances and variations present within the educational data from Jazan University's College of Business Administration. The elaborate structure implies that the model is equipped to handle a wide array of inputs, suggesting a tailored approach to predicting student performances. While extensive trees can sometimes be prone to overfitting, the perfect prediction accuracy reported previously gives an optimistic outlook on the model's real-world applicability. It may well be that the model has struck an impressive balance between learning detailed patterns and maintaining the ability to generalize to new data. Moreover, the level of detail captured here could provide invaluable insights into the factors affecting student success, allowing for more nuanced interventions and support tailored to individual needs. In essence, this decision tree is not just a model but a map of discerning educational insights, with the potential to guide policy makers and educators towards informed decisions that can positively influence the future of academic success at Jazan University.

## VI. Conclusion

This work presented a comprehensive investigation into the development and evaluation of the Predictive Learning Analytics Model (PLAM). The systematic approach encompassed research design, data collection, model development, and evaluation, aiming to enhance Teaching and Learning Outcomes (TLOs) by leveraging predictive analytics and machine learning techniques. Starting with the theoretical foundations of predictive learning analytics, key concepts, methodologies, and applications in the field were elucidated. A conceptual framework, drawing upon educational theory and data science principles, guided the research and informed the design and implementation of the PLAM. The empirical phase involved data collection from CBA students, including demographic information, academic performance metrics, and engagement indicators. This dataset formed the foundation for constructing and training the PLAM, using various classification algorithms to predict student outcomes and identify actionable insights for instructional improvement. Rigorous model evaluation and performance analysis assessed the effectiveness of the PLAM in predicting student performance and informing instructional strategies. Performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC provided insights into each algorithm's strengths and limitations, promoting evidence-based decision-making and continuous improvement in educational outcomes. Data prediction and model limitations were also explored, identifying potential challenges and areas for future research and development. The Attribute-Relation File Format (ARFF) facilitated seamless data processing and analysis within the PLAM framework. Additionally, the characteristics of the Random Tree model were investigated, highlighting its unique attributes and potential implications for educational practice. Overall, this work contributes to the growing body of literature on predictive learning analytics and its applications in higher education. By leveraging advanced data analytics techniques and machine learning algorithms, the aim was to enhance instructional effectiveness, promote student success, and drive continuous improvement in educational outcomes at CBA and beyond. Future research may include refining the PLAM, integrating additional data sources and predictive features, and exploring novel machine learning approaches to address complex educational challenges. Sustained collaboration and innovation can harness the power of

predictive learning analytics to transform teaching and learning practices, fostering a more inclusive, equitable, and effective educational environment for all stakeholders.

## References

[1] S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for Educational Data Mining: A Review," J. Educ. Behav. Stat., vol. 42, no. 1, pp. 85–106, Feb. 2017, doi: 10.3102/1076998616666808.

[2] L. Ji, X. Zhang, and L. Zhang, "Research on the Algorithm of Education Data Mining Based on Big Data," in 2020 IEEE 2nd International Conference on Computer Science and Educational Informatization (CSEI), Jun. 2020, pp. 344–350. doi: 10.1109/CSEI50228.2020.9142529.

[3] K. Wongsuphasawat, Y. Liu, and J. Heer, "Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study." arXiv, Nov. 01, 2019. doi: 10.48550/arXiv.1911.00568.

[4] M. Staniak and P. Biecek, "The Landscape of R Packages for Automated Exploratory Data Analysis," R J., vol. 11, no. 2, p. 347, 2019, doi: 10.32614/RJ-2019-033.

[5] C. Baek and T. Doleck, "Educational Data Mining versus Learning Analytics: A Review of Publications From 2015 to 2019," Interact. Learn. Environ., vol. 0, no. 0, pp. 1–23, Jun. 2021, doi: 10.1080/10494820.2021.1943689.

[6] A. Nguyen, L. Gardner, and D. Sheridan, "Data Analytics in Higher Education: An Integrated View," J. Inf. Syst. Educ., vol. 31, no. 1, p. 61, Mar. 2020.

[7] G. Casalino, G. Castellano, and G. Vessio, "Exploiting Time in Adaptive Learning from Educational Data," in Bridges and Mediation in Higher Distance Education, L. S. Agrati, D. Burgos, P. Ducange, P. Limone, L. Perla, P. Picerno, P. Raviolo, and C. M. Stracke, Eds., in Communications in Computer and Information Science. Cham: Springer International Publishing, 2021, pp. 3–16. doi: 10.1007/978-3-030-67435-9_1.

[8] M. Klose, V. Desai, Y. Song, and E. Gehringer, "EDM and Privacy: Ethics and Legalities of Data Collection, Usage, and Storage," International Educational Data Mining Society, Jul. 2020. Accessed: Apr. 08, 2023. [Online]. Available: https://eric.ed.gov/?id=ED607820

[9] A. A. Mubarak, H. Cao, and S. A. M. Ahmed, "Predictive learning analytics using deep learning model in MOOCs' courses videos," Educ. Inf. Technol., vol. 26, no. 1, pp. 371–392, Jan. 2021, doi: 10.1007/s10639-020-10273-6.

[10] A. Namoun and A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," Appl. Sci., vol. 11, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/app11010237.

[11] S. Ranjeeth, T. P. Latchoumi, and P. V. Paul, "A Survey on Predictive Models of Learning Analytics," Procedia Comput. Sci., vol. 167, pp. 37–46, Jan. 2020, doi: 10.1016/j.procs.2020.03.180.

[12] Y. Cui, F. Chen, A. Shiri, and Y. Fan, "Predictive analytic models of student success in higher education: A review of methodology," Inf. Learn. Sci., vol. 120, no. 3/4, pp. 208–227, Jan. 2019, doi: 10.1108/ILS-10-2018-0104.

[13] "Open Learning Initiative," OLI. https://oli.cmu.edu/

[14] "Learning to grow machine-learning models," MIT News | Massachusetts Institute of Technology, Mar. 22, 2023. https://news.mit.edu/2023/new-technique-machine-learning-models-0322

[15] "About the CVC-OEI – California Virtual Campus." https://cvc.edu/about-the-oei/

[16] "Harvard University | edX." https://www.edx.org/school/harvardx

[17] S. Hussain, R. Atallah, A. Kamsin, and J. Hazarika, "Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study," in Cybernetics and Algorithms in Intelligent Systems, R. Silhavy, Ed., in Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2019, pp. 196–211. doi: 10.1007/978-3-319-91192-2_21.

[18] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," Telemat. Inform., vol. 37, pp. 13–49, Apr. 2019, doi: 10.1016/j.tele.2019.01.007.
[19] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," Comput. Hum. Behav., vol. 104, p. 106189, Mar. 2020, doi: 10.1016/j.chb.2019.106189.
[20] M. Salihoun, "State of Art of Data Mining and Learning Analytics Tools in Higher Education," Int. J. Emerg. Technol. Learn. IJET, vol. 15, no. 21, pp. 58–76, Nov. 2020.