Early Identification of Vulnerable Students with Machine Learning Algorithms

MOHAMMAD HAMZA AWEDH[1], AHMED MUEEN[2]
[1]DEPARTMENT of Electrical and Computer Engineering,
Faculty of Engineering, King Abdulaziz University,
SAUDI ARABIA

[2]DEPARTMENT of Computer and Information Technology, King Abdulaziz University,
SAUDI ARABIA

**Corresponding Author*

Abstract: - Education is an important component in defining the overall development of a country. It is also a significant tool for achieving success in life. One of the important aspects influencing any educational institution's success is its students' academic achievement. In educational institutions, student dropout is a complex problem. Educational managers consider it vital to predict a student's risk of dropping out as soon as possible. It still needs to be easier to predict accurately in advance. The major problems in the present research work include overfitting in a predictive model, complex variable relationships, insufficient feature extraction, and data pre-processing complexity. The key goal of this study is to improve student achievement, decrease the number of dropouts, create support plans, and constantly modify these plans based on ongoing progress monitoring. Specifically, this research aims to identify at-risk students early using machine learning algorithms, allowing educational institutions to take timely and targeted interventions. Identifying the student's needs early in their time with you will ensure that vulnerable students get the support they need, help prevent dropout rates from increasing, and significantly benefit their general academic performance. In this work, the King Abdulaziz University database was used. Exploratory Data Analysis (EDA) is heavenly for understanding the characteristics of the data, identifying anomalies, recognizing trends, and directing further data pre-treatment procedures. Genetic Algorithm-optimized Latent Dirichlet Allocation (GA-LDA) is used for feature extraction. We utilize canopy clustering with a Gaussian Flow Optimizer (GFO) for accurate student grouping. Finally, a hybrid Logistic Regression-K-Nearest Neighbour (LR-KNN) technique is used for data classification. Accuracy, precision, recall, F1-score, sensitivity, and specificity metrics were used to examine the proposed model.

Key-Words: - Machine learning, Data mining, Feature extraction, Data classification, Gaussian Flow Optimizer, Regression-K-Nearest Neighbour, Exploratory Data Analysis, Education.

Received: May 2, 2024. Revised: November 24, 2024. Accepted: December 27, 2024. Published: January 27, 2025.

1 Introduction

In the fast-remodeling world of education, there is a wide search for better modes of teaching and learning. The passports of graduates and postgraduate students are coming out in large numbers every year, and the growth of educational institutions has also seen a rise recently. It has been established that student achievement is one of the most critical matters to educational managers, [1]. In the present global scenario, academic institutions have started relying on data solutions to enhance the quality of the teaching-learning process and students' achievements. This commitment to early tracking of students likely to experience poor academic performance is especially important to ensure that such students receive the educational support they require to excel. Scholars from fields such as higher education are beginning to actively participate in gaining more sophisticated information from data sources. A recent merging and alliance have evoked the study field combining education with Data Mining (DM), [2]. College registration has increased

due to the recent diversification of colleges into new areas of learning. The systems of higher education across the globe have developed rapidly during the last fifty years. Hence, more youths are in higher education institutions today, and old students have more chances, [3]. Because of increased enrollment, large quantities of data have been generated, which, if analyzed effectively, can create valuable information. Inconsistent data formats and insufficient evaluation tools and methods contribute to adequate decision-making, [2], [4]. Universities need to learn how not to lose their students if they are to perform well. The student dropout rates are an important indicator of the delivery of services, in this case, the quality of education. The issue of student attrition is multifaceted in the educational system, and it has positive effects on students and negative impacts on institutions, society, and the economy, [5]. It is a massive problem for the countries with the 'American dream' economy and the developing states. A dropout is a very negative thing for a student, and it has a negative financial impact on the university, [6]. It was noted that many managers of the educational system experienced the challenge of predicting the probability of leaving college as soon as possible. It is increasing and getting more prevalent; however, this is one of the most challenging issues in early prognosis. While it is a recent frequency on the rise, early prediction is not easy, [7]. DM is a strong method that enables extracting relevant data from large data sets. To enhance student grades and decrease the weak students' failure rate, DM can observe this poor behavior correctly, [8]. A university can professionally distribute resources by predicting the likelihood of students for specific courses, assisting learners in making wise choices of their fields of specialization, courses, and programs, and assisting them in boosting their performance according to data collected in the past. This implies that the university policy shall be forced to address the string issues, probably directing most of its concern towards student support systems, [9]. Data mining is an efficient tool for providing information that is otherwise unavailable and useful for upcoming predictions and suggestions in the education sector. Applying data mining to the attention behaviors of students in long lectures with active computer usage, it is possible to distinguish different behaviors, [10]. When used in the context of studying the factors/her factors that affect students' performance, it employs several techniques/algorithms to learn from the diverse data feeds. DM can only be realized by abstracting simple data from a stored data set that has not been discovered, [11]. The previous studies support this theory and note that it is good that students are aware when they make a wrong decision or that other known and unknown factors affect the decision. This is why it is important to minimize the dropout rate of students and identify the factors that can potentially influence the student's decision to drop out of school. It is also said that the educational system issue, which is perceived to be one of the hardest, is that of students dropping out, [12]. Since university dropouts also feel failed, they are detrimental in terms of the economy and society as well as to the self-image of the learners. The development of models to estimate the probability of non-continuation in higher education and to define students with a high risk of withdrawal to apply the retaining rules is the issue that attracted much attention to researchers, [13]. For the institutions to be able to make selections that would produce the best outcomes, they must have access to all education data in the various formats they come from various sources, [14]. For enhancing graduation rates in higher education, it is required to improve student retention, [15]. Moreover, given retention rates provide information about the opportunities, including planned student counseling. The primary goal is to discuss what constitutes the elements and how one can build the model that optimizes the predictability for using data mining techniques in the early identification of students who may not be successful in their careers.

For this study, we employed data from the King Abdulaziz University archive, which consists of students' performance, attendance, and behavior data, which are important for the early identification of at-risk students. Exploratory Data Analysis (EDA) is applied to identify data characteristics and detect or get to know the data patterns and subsequent EDA data pre-processing steps, which may include missing values, outliers, and normalization. The proposed Latent Dirichlet Allocation algorithm is optimized by a method such as Genetic Algorithm (GA) to enhance feature extraction and determine multiple patterns related to student risk. To discover relevant features, the study uses canopy clustering with GFO to accurately group students. It also uses a combination of the Logistic Regression-K- Nearest Neighbour to enhance classification correctness.

1.1 Motivation and Objectives

The overall problem statement has several significant issues that need innovative solutions when compared to current approaches.

- **Overfitting in a Predictive Model**: In existing methods, the danger of overfitting when developing predictive models, when a framework fits sample data unusually, yet its dependability is bargained because it cannot generalize to new, unexpected conditions.
- **Complex variable relationships**: The existing method does not efficiently capture the deep associations that exist between the variables, which could reduce the correctness of the model in predicting students who are at risk of dropping out.
- **Insufficient Feature extraction**: The effectiveness of predictive models deeply depends on feature extraction. In current methods, inadequate feature extraction may result in models struggling to capture complex patterns, particularly in high-dimensional or difficult data.
- **Data Pre-processing Complexity**: In present methods, the early stages of examination, handling outliers, missing values, and deciding on the need for data pre-processing can be tough. These tasks may delay the analysis process and affect the quality of knowledge of insights.

This work is motivated by a desire to apply data mining to improve the educational process in our universities in Saudi Arabia. Goal one proposes that early intervention, informed by students 'performance data, can cut dropout rates and improve academic achievement.

The specific objectives are as follows:

- To identify patterns, correlations, and problems in student behavior and performance; get a holistic understanding of data.
- To detect missing values, outlier values, and variable encoding to guarantee disaggregated data reliability before analysis.
- To extract relevant features from raw data and construct predictive models to consider key factors predicting student success.
- Cluster algorithms are used to group students of similar characteristics together while, at the same time, classification models can be employed to produce predictions for students at risk.
- To implement early interventions, including academic support and counseling, based on analysis results, aiming to improve overall student achievement.

1.2 Research Contributions

This comprehensive research project encompasses the following key highlights:

- The data is obtained from the King Abdulaziz University database, which contains students' academic performance records, attendance, and behavioral data. These are indispensable data sources that can help to discover students who are at risk.
- EDA helps in understanding what attributes are present in the data and the existence and identification of outliers and patterns. It leads to subsequent stages of data pre-processing. There is also standardization and normalization of data before analysis and handling of missing values and outliers, among others.
- To extract features for the identification of 'hidden trends' or patterns of student risks, what is referred to as the GA-LDA is used.
- The research interprets model decisions to determine which characteristics are potent for decision-making. It uses the canopy cluster with the Gaussian Flow Optimizer to classify students with high accuracy. Also, the method used for classification combines both the Logistic Regression-K-Nearest Neighbour (LR-KNN) classification.
- The underlying goals of the initiative are to provide the target student's specific individual learning plan with adequate academic support, daily progress monitoring, and appropriate changes in academic support.

1.3 Paper Organizations

The study's remaining sections are organized as follows: existing research shortcomings are evaluated in Section II. Section III describes the specific problem statement. The proposed work is explained fully in Section IV along with diagrams and pseudocode. The experimental setup is provided in

Section V along with thorough explanations of the simulation setup, comparative analysis, and research summary. Section VI provides a thorough discussion of the suggested works' conclusion and future undertakings.

2 Literature Survey

Besides this section also elucidates in detail the limitations of prior research. This particular study, [16], aims to examine the relationship between students' first-semester academic productivity, particularly the first test and assignment, and their overall performance after a semester. The study's findings pertain to the chosen courses, which are probably not generalized to other fields of knowledge and different types of courses. This research [17] aims to use ML and education data mining to predict the final exam results of undergraduate students based on their midterm test scores. The classification accuracy achieved by the study is mind-blowing; however, there is an obvious failure to capture the variations of the ranges of student performance, for instance, differentiate between high-risk failure students and those who require moderately low assistance. The present study [18] will design a new rule-based system called Risk Flag (RF) to improve student learning achievements, as well as the performance of educational facilities. However, this approach pays much attention to an initial evaluation of students while disregarding other factors affecting students' performance, and the effectiveness of interventions also depends on individual student conditions. This study [19] looks at the problem of predicting student dropout rates from e-learning courses using data available in educational systems. Complex handling of huge amounts of educational data may cause the use of computer resources and may extend processing time, especially when analyzing data that extend to several years. Authors in [20] explore the stochastic ML to the challenge of predicting, from different times in a school year, which students are likely to fail future examinations. Probabilistic ML models are normally complex, much more so than deterministic ones, and their use implies the need for large computations, which is a problem of scale.

This paper looks at the attrition rates of universities and their implications on academia and the economy, [13]. Multiple methods of feature selection and multiple machine learning models used as a single approach are applied in the study. To end up with a useful propositional-level prognostic system for dropout rates, significant and highly capable computing resources, and extensive and intensive specialist training are required. This paper proposes an RNN-based network structure to solve the issue of identifying which online courses students are likely to need to complete. When essential data is missing or lossy, the RNN model will perform poorly; its performance depends on the quality and quantity of available learning features. This paper [23], examines the effectiveness of a structured intervention program aimed at enhancing the comprehension and study skills of Latin American undergraduate medical students. The efficacy of processes varies based on the institutional setting and student population; thus, the findings of this research couldn't be applicable to other medical school scenarios or student populations. This study [24], focuses on assessing secondary school students' academic performance since it has a big influence on their future educational results. The research is limited by the quality and availability of the anticipated data since inaccurate or missing data can lead the conclusions to be less accurate. This study [25], examines the challenges that online learning settings provide for students who are at risk of experiencing difficulty with concentration and academic performance. An approach for forecasting which students may be at risk is to study how they learn and deploy Deep Learning (DL) and ML algorithms. It depends on the speed and reliability of data input to continuously monitor and intervene. In real-time operations, model execution consumes a lot of processing power.

This study [26], uses ML approaches to identify whether students are at risk inside a learning environment. It tries to establish a mixed-model preliminary forecast system of student performance using various ML techniques. Many features in the data had to be represented in different ways. For example, this would hurt the model's generalization and understanding compared to the errors seen in error computation alone. A paper [27], in which the author presents his ideas on how to change existing practices of evaluation to something new: for example, not all examinations that mostly determine whether one passes or fails to be replaced by one that looks at fundamental abilities, colleges, and universities nowadays enter to admit students; it is not simply enough passing examinations as they did

before. It is technically difficult to integrate predictive algorithms into current protocols and educational systems; thus, platform and database compatibility is required. The primary focus of this study [28], is the early detection of university dropout risk, accounting for variables including academic exhaustion, satisfaction, and intention to discontinue education. The research creates a screening tool, but it does not provide an in-depth evaluation of how well it predicts which students will drop out in the end. In this study [29], DM and ML to forecast student retention at various higher education levels. Perhaps the problem of overfitting has not been effectively solved, hence lowering the model's robustness; this is where models are built to perform well on the training data but poorly on new data. During the COVID-19 lockout, the study [30] developed the UBU Monitor tool, which helped the institution determine vulnerable students or those who might drop the health science degree program. The study's execution of the UBU Monitor instrument in a specific topic limited the assessment of its effectiveness across all academic courses and disciplines. In this study [31], a novel prediction approach was created to fully use this learning behavior data. Our method creates a hybrid deep learning model that can concurrently extract the temporal behavior information and the overall behavior information from the learning behavior data, helping to more accurately predict the high-risk students. The implementation cost of these models rises due to their higher computing demands, particularly for educational institutions with tight budgets. This article [32], suggests a method for automatically monitoring and predicting student grades and marks. This study aims to minimize RMSE and enhance classifiers' efficiency. The data set in the proposed method is preprocessed for data cleaning, whereby only the student's labeled academic history data is used to train the regression model and the DT classifier. In the case of time series data, which is helpful in an educational environment to track students' progress over a given period, RF is not the most suitable. In this study [33], the authors employed DM and video learning analytics resources to estimate the students' performance for the semester. Seven machine learning classification algorithms: Naïve Bayes, K- Nearest Neighbour (KNN), Logistic Regression (LR), Artificial Neural Network, Support Vector Machine (SVM), Random Forest (RF), and the Decision Trees (DT) were employed to analyze data from mobile applications and the LMS and SIS platforms. However, KNN is not very useful in determining the variables that impact student performance as it doesn't reveal the relationship between features and the target variable. This study offers a special approach for predicting student performance [34], which involves ML techniques. An example of augmenting feature selection is using the k-means algorithm with the EM and genetic algorithms.

Further, they employed a set of ML classifiers to examine the recommended findings. Since Naive Bayes uses a simplistic representation, it fails to capture complex patterns in the data set. If the model does not capture equity in the various aspects of student performance, it produces underfitting. Table 1 (Appendix) describes the research gap in the literature survey.

3 Problem Statement

An ongoing key problem is the improvement of the efficiency of educational procedures in Saudi universities. A few of the issues highlighted in the latest research are as follows: An ongoing key problem is improving the efficiency of educational procedures in Saudi universities. A few of the specific issues highlighted in the latest research are as follows:

This paper [35] considers assessing the use of DL in EDM for predicting students' performance and their ability to identify low-performing students in courses like 'Programming' and 'Data Structures.' Therefore, using a dataset from a four-year university, several models such as Decision Trees (DT), Logistic Regression (LR), gradient boosting, Random Forest (RF), K-nearest neighbors (K-NN), and support vector classifiers are built. This study also encompasses other areas, such as predictive models and bias analysis, and the educational authority has an API that can interface the results of the models. Some of the problems detected in these papers are: Some of the issues detected in these papers are:

• Indeed, feature transformation and selection are informative parts of the entire progress of the model. Many things can be said about the confusion when identifying the most important aspect from a list of parameters.

• topics like aspects of importance in data, detecting important values, and handling with no

results are common problems encountered when engaging in the first levels of data analysis.

• This means that when there are no clustering algorithms, certain targeting and intervention chances may not be noticed and, hence, not administered. Implementing an intervention program requires grouping to increase the likelihood of success in the learning process.

• As the data volumes rise, the scalability problem during data preparation leads to delays in risk evaluation and instructional learning analytics.

The study [36], covers data from over 258,000 students between 2015 and 2020, employed learning analytics and machine learning models to predict at-risk students. It focuses on identifying students at risk of academic failure or dropping out in Uruguay's secondary education system. The paper designs predictive models, using the random forest algorithm, to operate at different times during the school year. The authors achieved high performance metrics, with AUROC scores exceeding 0.90 and F1-Macro values over 0.88, indicating strong accuracy in identifying at-risk students.

To improve early warning systems in educational institutions, this paper [37], suggests an enhanced fuzzy clustering method based on composite components. The purpose of smart campuses is to help students make decisions and make sure they can complete their education. This work [38], addresses the issue of figuring out academic accomplishment for students at Higher Education Institutions (HEI) using data-mining approaches. The objective is to identify features that group to represent different student performance levels and to develop prediction models for each cluster to fully detect student performance. Problems faced by researchers are:

• Poor feature extraction reduces the accuracy of the predictions as it cannot capture the proper patterns and information of the data. This stage is sometimes overlooked when dealing with complex or 'Big Data', but when this occurs, models fail to give the right prediction.

• Higher computational cost and demand for resources with small improvements of enhanced fuzzy clustering technique make them unsuitable for large problems or real-time applications.

• K-means reflects data points into a single cluster, which is problematic if the data point's cluster membership is not well defined. The clustering is made by probabilistic cluster assignments are generated, and soft clustering algorithms generate the result.

• SVMs need precise hyperparameter modification and perform poorly on huge datasets. Furthermore, they could have trouble with complicated non-linear connections or noisy data, which leads to inadequate prediction results.

The complicated problem of student-teacher attrition is examined in this article [7], which has major consequences for people, organizations that educate people, and society. To improve attrition prediction and early identification of student teachers who are in danger, this effort aims to develop a predictive algorithm. Some of the major problems in this study are:

- Overfitting the data during the construction of a predictive model results in the model performing well on sample data but having difficulty generalizing to new, unexpected scenarios.
- The research's four-step logistic regression technique could not have captured complex correlations between variables, which limits the model's capacity to forecast future events with any degree of accuracy.

3.1 Research Solutions

One strategy for this could be models that use featured analysis for identifying the students who are in danger of becoming dropouts based on data Similar to the ones on attendance, students' behavior, and their performance. Lessons learned: This model aims to minimize fallbacks and dropouts, seen as unnecessary expenses, by accelerating the period to identify at-risk pupils and provide support solutions via an elongated ML range. The information used in this work is the data from King Abdulaziz University, which contains students' behavioral, attendance, and academic performance information. That is why these data sources significantly help define which students are at risk. Based on that definition, the task of EDA is to show or prove what information may be obtained from the contrast of data among different systems and retain information on specific aspects. As is true with any hypnotic suggestion, it leads to other data processing operations. Namely, you must learn missing numbers and outliers during the preprocessing (normalization) phase to analyze it. Thus, GA-LDA facilitates the feature extraction and the identification of more elaborate student risk profiles. Thus, by applying the

canopy clustering alongside the Gaussian Flow Optimizer, the study group improves the exactness of the clusters and the interpretation of model choices to reveal the relevant characteristics. Furthermore, it incorporates both the LR and KNN algorithms for classification. This entails developing client-specific treatment and implementation, continuous assessment of the learner's performance, and constant adjustment of the approaches to enhance interventions to help at-risk teenagers learn.

4 Proposed Method

However, we must identify the key signs of the students likely to drop out through the proposed ML algorithm, which includes performance, attendance, behavior, and other aspects. This strategy is designed to keep the learners engaged and minimize dropout rates by identifying helpless learners early and offering it through different ML approaches. It improves the educational outcomes and makes the learning environment more positive. The proposed architecture is illustrated in Figure 1 (Appendix). The following are the main steps in this approach.

- Data collection
- Exploratory Data Analysis (EDA)
- Data pre-processing
- Feature extraction
- Interpretation and evaluation
  - Clustering the data
  - Classify the data
- Early intervention and support

4.1 Data Collection

Our primary study data source will be the King Abdulaziz University (KAU) database, which contains behavioral information that helps identify at-risk students as well as academic achievement and attendance records.

4.1.1 King Abdulaziz University Database

This database contains the behavioral, academic achievement, and attendance records of the students. Gender, nationality, place of birth, state ID, grade ID, section ID, topic, semester, relation, raised hands, visited resources, announcement's view, discussion, parent answering survey, parents school satisfaction, student absence days, class are the records present in KAU database.

4.2 Exploratory Data Analysis (EDA)

EDA is a preliminary data analysis process in which data is assessed and depicted to gain preliminary insight into properties of values, including outliers and patterns. Thus, EDA aids an analyst in deciding which variables require further analysis and which require pre-processing. Histogram: This is also a graphical display, but the special feature of this data representation tool in EDA is that it represents a distribution of a given variable. Histograms reveal central and distribution characteristics, breaking data points into certain ranges. Histograms have a vital role in revealing the basic properties of the data and help in making the proper decisions concerning the further stages of the analysis.

Exploratory data analysis, or EDA, comprises many basic parts that help to understand and get insights into the data or structure of the data set.

- Descriptive Statistics: Descriptive statistics gives an accurate measurement that summarizes the main characteristics of the dataset that it contains:
  - Mean: The data set's average can be obtained by summing all the values and dividing them by an overall count of values.
  - Median: The middle value of the data when this data has been arranged either in descending or ascending manner. It is divided into two equal parts.
  - Mode: This, of course, can be challenging, so more often than not, the value or values that rank first are the most frequently observed.
  - Standard deviation: The difference between the values and the mean. It defines the spread of the data values in the analysis.
  - Variance: The average value subtracted from the squared values of the measured values and the result divided by measurement range. Knowing how the data is distributed is useful, and this provides for the distribution of data.
- Data Visualization: These methods are nevertheless necessary for one to see the information as the analysts work so that one can easily grasp it. Namely, numerical data are displayed in histograms, box and scatter plots, bar and heat plots only, and only frequencies, tendencies, outliers, and variable dependencies are discussed.
- Data Cleaning and Preprocessing: Some processes that occur before data preprocessing are as follows: Data cleansing involves identifying data duplicates, errant data points,

missing values, and inconsistent data within the existing database. It will be typical for data preparation tasks before other data analyses, including controlling some of the categorical variables, feature scaling, and data standardizations.

- Feature engineering is a process that entails adding another set of features to another set of features. It is called feature engineering, as it aims to redefine a given dataset to enhance performance in a specific forecast or explanation.

4.2.1 Histogram

Histograms refer to a process in which continuous data information or a set of data is represented. They divide observations into bins or classes according to their value. Therefore, since they split one variable into intervals or bins and represent the rates or counts of observations falling in each bin in the form of bars, histograms are used to show the distribution of that variable.

4.3 Data Pre-Processing

Pre-processing from data collected through EDA includes managing missing points, outliers, and data scaling.

4.3.1 Handling Missing Values

This data management approach involves using a means to enter the missing values through the mean of the specific input. That method concerns variables that are normally distributed.

4.3.2 Addressing Outliers

A logarithmic transformation decreases the impact of outliers and reduces the range of values in a dataset. This technique has a certain advantage when the data distribution is positively skewed.

4.3.3 Normalization

Normalization, one of the data pre-processing techniques, is used to bring the scale of the data to an acceptable level for comparison. Quantile transformation transforms the data to align them with specific probability distribution functions, such as the normal distribution.

As discussed in the previous section, quantile transformation can help handle time-series outliers and must be applied to the input time series. Quantile transformation is a statistical transformation method that aims to make the distribution of data points in a dataset uniform or normal, hence reducing the influence of outliers. In several ML and data analysis applications, obtaining more consistent and reliable results to achieve this is useful.

4.4 Feature Extraction

Currently, the Genetic Algorithm-optimized Latent Dirichlet Allocation, or GA-LDA, expands the feature extraction approach application in the risk and educational analysis fields. In GA-LDA, conversely, hyperparameters are optimized using genetic algorithms, making a neat fit on data in contrast to the fixed hyperparameters, which is common in the conventional LDA approach. It is especially useful in identifying variable themes in textual data such as student behavior logs, course content, and assessment. It thus improves the transformation of the document into document topic distribution. It noted that compared with other methods, the importance of LDA features is rather high for GA-optimized in capturing different patterns concerning student risk, which is very important in the context of prediction. These factors are used because they show different risk profiles concerning students, which gives assurance to them as input to any subsequent predicting or classifying task that seeks to establish whether or not a student is at risk.

4.4.1 Genetic Algorithm (GA)

One among a robust class of evolutionary algorithms for solving search-based problems that can be implemented in many searching problems is the Genetic Algorithm (GA). This is because GA does not search for other various random solutions like random local search algorithms but tries to keep a record of the best solutions. Algorithm 1 shows the major activities involved in the GA's working principle.

Algorithm 1: GA working mechanism
Step 1: Initiate a random population of individual solutions.
Step 2: Analyze each solution's fitness solution.
Step 3: Selection of the top individuals.
Step 4: Utilize crossover.
Step 5: Employ mutation.
Step 6: Generate a new population.
Step 7: Repeat step (2) to step (6) until convergence or set of iterations.

Step 1: Initializing a group of solutions (individuals) is how the GA gets started. They are all referred to as the population. Naturally, every solution is referred to as a chromosome that contains genes. Random initialization is applied to the solutions inside the lower and higher bounds. To convert the random initiated values to the alternatives, use equation (1), where the list is the list of alternatives and rand denotes the random value that was established. The item's index in the options list is the result.

$$Index = [rand * (length(list) - 1)] \quad (1)$$

Step 2: Every solution's fitness function is computed. By the Survivor Selection Policy, a fitness function is employed to assess a solution's fitness for carrying over into the following generation. Weighted Sum (WS) value from equation (2) is used to compute the fitness function. Following the DL model's training and assessment of the whole dataset, the WS internal metrics are determined. Prior to proceeding to the next stage, the solutions are arranged according to the fitness value, first in ascending order for minimization issues and descending order for maximization problems. Importantly, age-based selection is an alternative to fitness-based selection. In this method, the newest chromosomes in the population are exchanged for the oldest.

$$WS = \frac{w_1}{Loss} + w_2 * Accuracy + w_3 * Precision + w_4 * Recall + w_5 * Specificity + w_6 * F1 + w_7 * AUC \quad (2)$$

where w_1, w_2,w_3,w_4,w_5,w_6,w_7 are the multiplied weights and their sum must equal to 1. The weight w_2, which belongs to the accuracy is set to be the highest among others.

Step 3: The selection process selects the best solutions that remain for the next generation. The applied task determines the number of best answers. It can be the top half of the top two answers.

Step 4: The crossover is significant in GA because it produces the offspring that will make up the population of the next generation. The crossover can be applied using a variety of techniques, including order-based, shuffle, ring, partial mapping, single-point, multi-point, uniform, and entire arithmetic recombination.

Step 5: The purpose of the mutation is to prevent early convergence and boost generational variety. A mutation is the application of the search space idea via a random alteration in the chromosome. A variety of mutation techniques exist, including random resetting, flipping a bit, swapping, scrambling, and inversion mutations.

Step 6: After successfully generating, the new generation has been set up to go on to the next iteration.

Step 7: Except for the initial step, every prior step is repeated by specified completion criteria. It might be several iterations or convergence, which would indicate that there are no appreciable differences between the current and past generations. It is possible to effectively arrive at the ideal or nearly ideal answer when the iterations are finished.

4.4.2 Latent Dirichlet Allocation (LDA)

A generative probabilistic model called LDA has been suggested to extract hidden concepts from different types of text sources. The words in text texts serve as the main data units for this unsupervised model. The smoothed LDA graphical model is shown in Figure 2. The $X_{d,p}$ the index of the word x in the document d, represents the input of the model. The model yields the M, which is a predetermined quantity of latent concepts. Each topic m, $m \in \{1, \dots \dots M\}$ is represented by the discrete probability distribution φ_k over the vocabulary V and generated from a Dirichlet distribution $\varphi_k \sim Dir(\beta)$. Additionally, every document d, $d \in \{1, \dots \dots D\}$ comes from a Dirichlet distribution $\theta_d \sim Dir(\alpha)$, which represents the distribution of topics for every document d. From θ_d we calculate $a_{d,p}$ per word topic assignment in the document d, where β and α are the Dirichlet parameters.

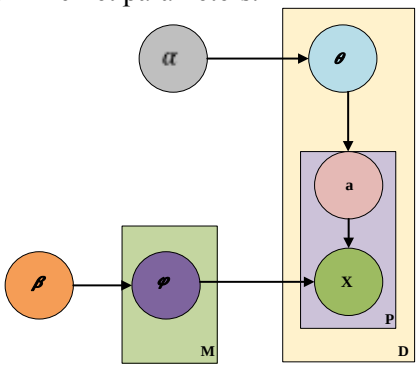

Fig. 2: Graphical model representation of LDA

Another method in enhancing feature extraction for the forecast of complimenting students' academic achievement is the Genetic Algorithm-optimized Latent Dirichlet Allocation, commonly known as GA-LDA. This method often scans through a list of possible indications employing a genetic algorithm to determine the most suitable ones. It is then followed by LDA Latent Dirichlet Allocation to detect latent ideas in the characteristics that have been chosen to give some meaning. GA-LDA uses both techniques to help find which aspects of a student's educational profile are the most useful. This helps arrive at informed decisions about the educational field and comes in handy in making correct estimations of the student's performance.

4.5 Interpretation and Evaluation

Interpret the model's decisions to understand which features are most influential in identifying at-risk students.

4.5.1 Clustering the Data

Utilize canopy clustering to organize students according to similar features. Canopy clustering is a hierarchical clustering algorithm with two main phases: Initialization and Clustering.

- *Initialization phase:* In the Initialization Phase of canopy clustering, we begin by determining the similarity criteria for classifying students into canopies by establishing T1 and T2 as distance thresholds.

4.5.1.1 Canopy Clustering

The data may be approximately divided into many overlapping groups, which are then recorded as Canopy using the Canopy algorithm. Every subset forms a cluster; to speed up clustering, low-cost similarity measures are often used. Consequently, Canopy clustering is often used for other clustering methods' initialization procedures. The initial data set X is sorted by specified constraints, and the creation of Canopy requires the specification of two distance thresholds: T1, T2, and T1 > T2. A rough distance computation method is used to determine the distance d between successive sample data vectors in A and X once a random data vector A in X is selected. The sample data vector with d less than T1 is mapped to a canopy, while the sample data vector with d less than T2 is eliminated from the available set of potential center vectors. Follow the preceding steps until either X is empty and the operation ends, or until there are no more possible center vectors on the list. Figure 3. demonstrates the procedure of canopy clustering. Here, using canopy clustering similar characteristics are grouped as cluster 1, cluster 2, and cluster 3 respectively.

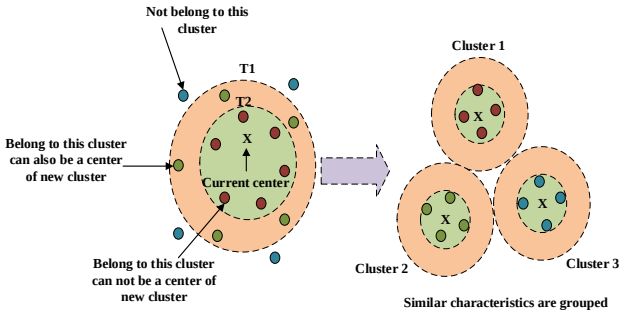

Fig. 3: The procedure of canopy clustering

T1 and T2 are tailored to our data and similar characteristics, such as using the Grade Point Average (GPA) or test scores to categorize academic performance. We begin by selecting at random the first students to serve as the initial center of the canopy. For each student in the dataset, we determine their distance from the current canopy center using an appropriate similarity metric, such as Euclidean distance for numerical features or a custom metric. If the distance is less than T1, the student joins the existing canopy. If the distance is less than T2, the current student replaces the center of the canopy. This procedure is repeated for each student, yielding a collection of canopies containing students with similar characteristics. These canopies serve as the basis for subsequent clustering based on student characteristics.

4.5.1.2 Euclidean Distance

Numerous clustering techniques use Euclidean distance metrics to identify clusters. It is assumed that m and n stand for the number of objects and characteristics and that there are x multidimensional spaces. When computing an interval using the Euclidean distance formula, the usual scaling formula is given as equation (3):

$$c(z_m, z_n) = \sqrt{\sum_{r=1}^{x}(z_{mr} - z_{nr})^2} \quad (3)$$

The standard scale of the interval in the Euclidean n-dimensional space model can also be determined using the cosine variable. This is the calculating formula shown in equation (4):

$$O_{cos}(z_m, z_n) = \frac{\sum_{r=1}^{x}(z_{mr} . z_{nr})}{\sum_{r=1}^{x} z_{mr}^2 \sum_{r=1}^{x} z_{nr}^2} \quad (4)$$

- *Clustering Phase:* In the Clustering Phase, we propose a Gaussian Flow Optimizer. Gaussian Flow represents the efficient and perfect modeling of data distributions using GMM, while Optimizer emphasizes the role of the Whale Optimization Algorithm (WOA) in fine-tuning clustering parameters. Utilize the Gaussian Mixture Model (GMM) to capture the distribution of academic performance characteristics within each canopy. Initialize the GMM parameters and establish a function for evaluating clustering quality.

4.5.2 Gaussian Mixture Model (GMM)

When processing non-spherical data sets, the GMM clustering technique provides more flexibility. The Gaussian probability density function was used to split the data set into many models; GMM uses the function to properly quantify a set of data. To do unsupervised clustering, the GMM probability model which also has the quickest learning speed—fits the input data set to create a suitable linear GMM distribution combination model. When reservoir composite quality is classified using GMM clustering algorithms, finishing quality and reserving quality within the same clustering category differ less, while finishing quality and reservoir composite quality within different clustering categories differ significantly. As a consequence, the reservoir composite quality index of the same splitting stage is equivalent.

- GMM clustering algorithms

The sample data set $Z' = \{z'_{11}, z'_{11} \ldots., z'_{xy}\}$ of a known reservoir composite quality index conforms to a K Gaussian distribution; then the GMM appears to be in the form of equation (5):

$$q(z|\theta) = \sum_{k=1}^{K} \alpha_k \mathcal{N}_k (z|\theta_k) \quad (5)$$

where $\theta_k = (\mu_k, \sum_k)$, and the unit Gaussian distribution $\mathcal{N}_k(z|\theta_k)$, which mean μ_k and covariance matrix $\sum_k$, is called a component of the GMM, where α_k is a mixed parameter, and is the weight of the k Gaussian distributions and represents prior probability:

$$\sum_{k=1}^{K} \alpha_k = 1, 0 \leq \alpha_k \leq 1 \quad (6)$$

The probability density function of $\mathcal{N}_k(z|\theta_k)$ is given in equation (7):

$$\mathcal{N}_k(z|\theta_k) = \frac{1}{(2\pi)^{d/2}|\sum_k|^{1/2^e}} -^{1/2[(z-\mu_k)^S \sum_k^{-1}(z-\mu_k)]} \quad (7)$$

Then, these GMM parameters will be enhanced with the help of an optimization technique called Whale Optimization Algorithm (WOA) for the maximum value of a fitness function, defined in terms of the negation of the objective function. The WOA optimization is performed iteratively over-optimized iterations to improve the GMM clustering. Moreover, one should consider certain metrics to assess the comprehensiveness of the clustering made. Expanding from this, look at students in terms of the district, examine the resulting clusters, assign students to these clusters, and get some insight into the academic performance pattern. This sort of technique uses the melting-pot shape of GMM to work well when the sections are of various forms and uses the optimization procedure of WOA to adjust the parameters of the segments to enhance the clustering of students with similar characteristics under each canopy.

4.5.3 Whale Optimization Algorithm (WOA)

Before presenting the WOA optimization ideas, it is vital to comprehend the three phases of optimization searching. The three primary phases of the algorithm searching for food, encircling its prey, and swimming spirally to feed replicate the hunting strategies of humpback whales. Figure 4 illustrates how a random probability factor $p \in rand\,[0,1]$ and a coefficient $|X|$ influence its selection.

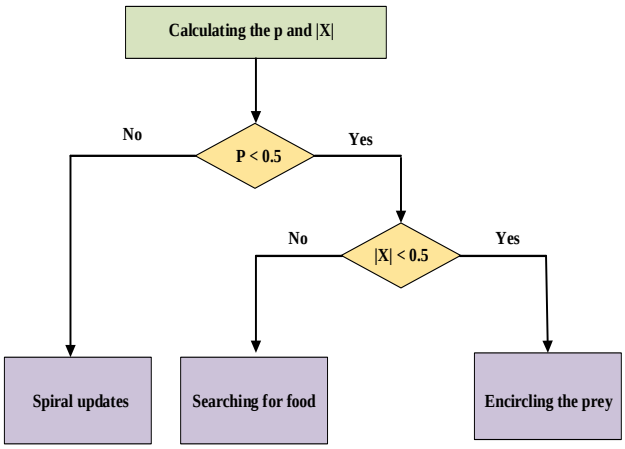

Fig. 4: Block diagram of WOA mechanism

4.5.3.1 Searching for Food

Whales search for food at random during the food search phase. The present individual whale chooses another individual whale at random to target and approaches its location. This procedure matches the

algorithm's worldwide development period. The following equation (8) should be referred to as:

$$\begin{cases} Z(s+1) = Z_{rand}(s) - B.C_1 \\ C_1 = |D.Z_{rand}(s) - Z(s)| \end{cases} \quad (8)$$

where $Z_{rand}(s)$ is a randomly selected individual whale from the existing whale population, $Z(s)$ is the current individual whale position, D is the vector of the coefficients randomly distributed between [0,2], and $|X|$ should be referred to as the following equation (9):

$$\begin{cases} X = 2x.r - x, \\ x = 2 - \frac{2s}{S_{max}} \end{cases} \quad (9)$$

where r is a random number within [0,1]. x is called the control parameter, s is the current number of iterations, and S_{max} is the maximum number of iterations. As the number of repetitions s grows, it can be seen that the value of the coefficient |X| similarly declines linearly from 2 to 0.

4.5.3.2 Encircling the Prey

Whale schools use bubble nets to assault their prey. It consists of two steps that correspond to the WOA algorithm's local exploitation stage: contraction and spiral updating. In WOA, the population member who has presently found the best solution is regarded as the target prey, and all other members of the population are drawn toward it. The following equation should be used to refer to the shrinkage envelope phase mathematical model:

$$\begin{cases} Z(s+1) = Z_{best}(s) - B.C_2 \\ C_2 = |D.Z_{best}(s) - Z(s)| \end{cases} \quad (10)$$

where $Z_{best}(s)$ is the best-positioned individual whale in the current population and C_2 is the length of the enclosing step. The smaller the value of $|X|$, the smaller the step length of the whale swimming.

4.5.3.3 Swimming Spirally to Feed

In the spiral renewal stage, other whales will hunt for food by swimming in a spiral toward the ideal whale. It causes them to look for the ideal person to work with to find the best potential answer. The location of the current whale is where the spiral update begins, and the position of the current best whale is where it finishes. The following formula is used to refer to the mathematical model.

$$\begin{cases} Z(s+1) = C_3 \,.\, e^{la} \,.\, cos(2\pi l) + Z_{best}(s) \\ C_3 = |Z_{best}(s) - Z(s)| \end{cases} \quad (11)$$

where C_3 is the distance between the current particular whale and the best-positioned whale, a is a constant coefficient and l belongs to a random value inside [0,1].

4.5.4 Classify Data

Implement classification algorithms for predicting at-risk students using the Logistic Regression with K-Nearest Neighbor (LR-KNN). Developing a hybrid model necessitates thorough consideration of the advantages and disadvantages of each component model. KNN can capture complex non-linear patterns, whereas logistic regression is useful for analyzing linear relationships. The hybrid approach seeks to improve predictive accuracy by utilizing both of these factors. Each subcluster's aggregated features are subjected to LR, effectively creating discrete models for students within distinct subgroups. This enables the identification of linear relationships within each subcluster, based on the assumption that students within the same subgroup share similar features and behaviors.

4.5.4.1 Logistic Regression

When using quantitative or qualitative independent variables to explain a binary response variable, statistical modeling is often used. It is a member of the generalized linear model's category. For instance, the usual form of the log odds for an LR model with a single independent variable, Z, which can be either binary or continuous, is $m = \beta_0 + \beta_1 z$, where the coefficients $\beta_0 \, and \, \beta_1$ are the regression parameters and z is the observed value of Z. This is non-linear model so that the odds are the exponent $e = a^{\beta_0 + \beta_1 z}$, representing a non-linear model as the chances are a non-linear mixture of independent variables, with the exponential function being assumed to be the base a. Then, the associated probability function of X=1 is

$$q = Q(X=1) = \frac{\exp(\beta_0 + \beta_1 z)}{(\exp(\beta_0 + \beta_1 z) + 1)} = \frac{1}{\left(1 + exp(-\beta_0 - \beta_1 z)\right)} \quad (12)$$

On the other hand, K-Nearest Neighbors is also individually applied to each subcluster to detect intricate non-linear patterns within these fine-grained student subgroups. Hybridization, the crucial concluding stage, combines the predictions from LR and KNN models within each subcluster. This fusion takes advantage of both the linear and nonlinear features of the models by employing a weighted average or other applicable techniques.

4.5.4.2 K- Nearest Neighbor

A crucial part of ML is the K-NN technique. It depends on the supervised learning methodology. All of the available data is preserved by the K-NN algorithm, which also classifies newly added data points according to how similar they are to previously classified data. It suggests that new information can be quickly classified into a well-defined category using the K-NN technique. The KNN technique is most often used for classification issues, while it can be used to regression as well. Newly collected data is categorized by the KNN algorithm into a group that closely matches the previously stored data from the training phase. Figure 5 shows the flowchart of KNN classifier. After grouping of similar characteristics using clustering is performed, K-NN is used to classify the clustered data as category A and category B.

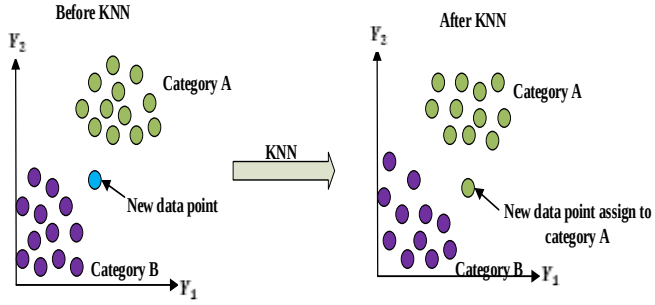

Fig. 5: Flowchart of K-NN classifier

High accuracy and precision in identifying at-risk pupils at every subcluster level are made possible by the hybrid LR-KNN architecture. In addition to offering a thorough knowledge of student behavior and improving the prediction accuracy in identifying students at-risk, it accommodates both linear and nonlinear patterns among different student groupings. A helpful ML technique for classifying student data based on academic accomplishment is LR-KNN. The method integrates the benefits of logistic regression, a powerful classification tool, with K-NN, a non-parametric methodology that considers the similarity of data points. Teachers and instructors can effectively place students into different performance groups by using this hybrid method. This enables them to provide targeted interventions and assistance to help students succeed academically.

4.6 Early Intervention and Support

To assist at-risk students in attaining academic success, perform assessments to identify academic and non-academic issues, establish individualized support plans, and continuously adjust those plans based on ongoing progress monitoring.

5 Experimental Results

This section presents the experimental analysis of the proposed DM approaches that are utilized to enhance the education of the students. The results demonstrate the suggested strategy's outstanding effectiveness. This sub-section includes the simulation setup, comparison analysis, and research summary.

5.1 Simulation Setup

The simulation environment is created to assess the feasibility of the proposed model and how well it predicts students at risk. The simulation tool used here is Python 3. 11. 3, and the system specifications are described in Table 2. The hardware requirements include a hard disk of 500GB, a minimum RAM of 2GB, and a processor of 2. 5GHz or higher. The simulation operates in a Microsoft Windows 10 (64-bit) environment. It also provides the best conditions for using computational methods since the hardware and software capabilities are sufficient for the successful functioning of the predictive model. The experiment follows this configuration, making the results replicable and scalable to similar educational prediction problems.

Table 2. System specification

Hardware specification	Hard disk	500GB
	RAM	minimum 2GB
Software specification	Simulation tools	Python – 3.11.3
	Processor	2.5 GHz and above
	OS	Windows 10-(64-bit)

5.2 Comparative Analysis

The proposed model is contrasted with other existing methods like fuzzy algorithm [37] and EnFftRP (Ensembled Fast Fourier Transformed Ranking Prediction), [21]. A number of metrics, including sensitivity, specificity, recall, F1-score, accuracy, and precision, are utilized to graphically evaluate the suggested model.

5.2.1 Accuracy

This corresponds to the classifier's capability and indicates the accuracy of the classifier. The fraction of accurate predictions divided by the total number of forecasts yields the accuracy. The following Equation (13) is used to compute it.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

where:

- True positives (TP): situations when a yes prediction is made.
- True negatives (TN): instances when no prediction is generated.
- False positives (FP): scenarios when the answer is really yes but the prediction is no.
- False negatives (FN): situations when a yes result is really projected to be a no.

Table 3. Numerical Outcomes of Accuracy (%)

No of Epochs	Accuracy (%)		
	Proposed	Fuzzy algorithm	EnFftRP
100	92	90	89
200	92.8	91.3	90.7
300	93.7	92.5	91
400	96.5	94	92.4
500	97.9	96.2	94.5

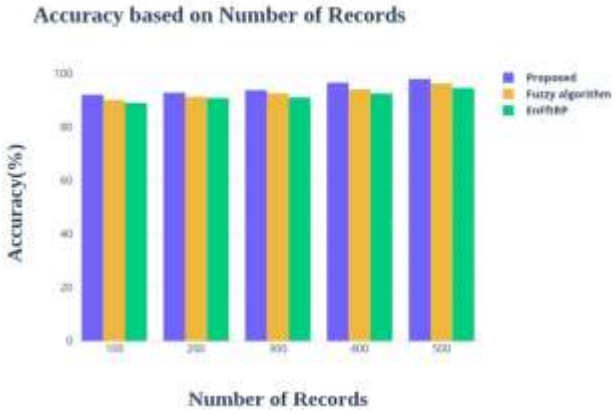

Fig. 6: Accuracy based on the Number of Records

Figure 6 and Table 3 showed a batch of three different algorithms used in the dataset. The methods described the implemented fuzzy algorithm presented in the paper and EnFftRP. The Proposed algorithm is more accurate during the epochs than others and increases from one hundred epochs from 92% to 97.9 percent in the subsequent 500 epochs. The Fuzzy Algorithm has an initial accuracy of 90% at the first 100 epochs and increases to 96.2% if the epoch number increases to 500, which shows reasonable improvement along with epochs but is slightly worse compared with the Proposed Algorithm and the accuracy factor. On the other hand, EnFftRP starts with 89% accuracy at 100 epochs and 94.5% at 500 epochs yet records the lowest accuracy rate between all the two models.

The number of epochs is relevant in identifying the learners who may require additional attention at an early stage using ML. The Proposed Algorithm is the most effective in that it guarantees finding vulnerable students early with maximum accuracy. Still, it can be assumed that the lower accuracy of the Fuzzy Algorithm in comparison with the Proposed Algorithm may mean its lower flexibility or optimality in terms of defining some students as vulnerable. When deciding between EnFftRP and a human language model, there tends to be a weakness in the specific prediction type: false positive or false negative, which could be considerable for its high-stake use cases, such as early students' intervention.

Meanwhile, the proposed algorithm is useful in scenarios where student data will be processed in real time to identify warning signs in the preliminary stages. Hence, by comparing these algorithms, educators and data scientists can select the model that best compromises accuracy and time.

5.2.2 Precision

The ratio of accurately categorized positive forecasts to all positively predicted forecasts, whether correctly or incorrectly classified, is used to measure precision. Equation (14) is used to compute it. Table 4 indicates the outcomes of precision.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (14)$$

Table 4. Numerical Outcomes of Precision (%)

No of Records	Precision (%)		
	Proposed	Fuzzy algorithm	EnFftRP
100	95	94.3	93.2
200	96.2	95	94.1
300	96.8	95.5	94.8
400	97.1	96.2	95.6
500	98.2	97.1	96.4

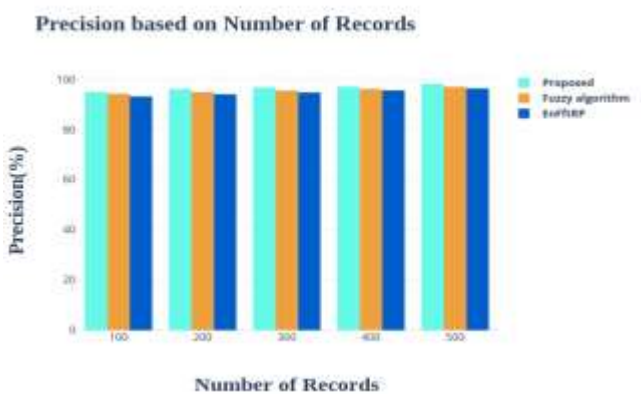

Fig. 7: Precision based on the Number of Records

Figure 7 shows how Accuracy is essential in classification models since it measures the prediction of positive results in a model. It calculates the ratio of true positive predictions to the total number of positive predictions the model makes, whether right or wrong. For all the sample sizes, it can be seen that the performance of the proposed algorithm is much higher than that of the Fuzzy Algorithm or EnFftRP in terms of the accuracy of its positive predictions. The Proposed algorithm has the highest Precision at all sample sizes, from 0.95 at 100 records to 0.982 at 500 records. This means the proposed algorithm outperforms the other two in accurately classifying positive samples. The Precision of the Fuzzy Algorithm is slightly below the proposed algorithm and above the EnFftRP algorithm, with the precision ranging from 94. 3-97. 1% with records added. This shows that for all the algorithms tested, there is a positive correlation between the number of records and the Precision achieved. This confirms earlier conclusions that harmonized and higher utilization works better when more information is given to an algorithm to work on. The Proposed algorithm presents results with precision ranging from 95% to 98.2%, therefore establishing that it stands to benefit from the availability of more data. The higher Precision of the proposed algorithm also means that it provides a better way to eliminate false positives compared to the Fuzzy Algorithm and EnFftRP, especially when false positives are very expensive or have negative effects like in medical diagnosis or firms' fraud detection.

5.2.3 Recall

Recall is computed as the ratio of correctly classified positive predictions to all positive predictions. Equation (15) is utilized to determine it.

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

Table 5. Numerical Outcomes of Recall (%)

No of Records	Recall (%)		
	Proposed	Fuzzy algorithm	EnFftRP
100	91.4	90	89
200	92.6	91.2	90
300	94.5	92.5	91.1
400	96	94	92
500	98	96	94.2

Comparing the proposed to the existing methods, such as 96% for the Fuzzy algorithm, 94.2 for EnFftRP, and 98% for the proposed method have higher recall than the existing techniques. Figure 8 and Table 5 indicate the outcomes of Recall. Higher recall metrics in predicting academic achievement show that the model is more successful in identifying a greater percentage of students who will do well in the educational setting, reducing the possibility of false negative results.

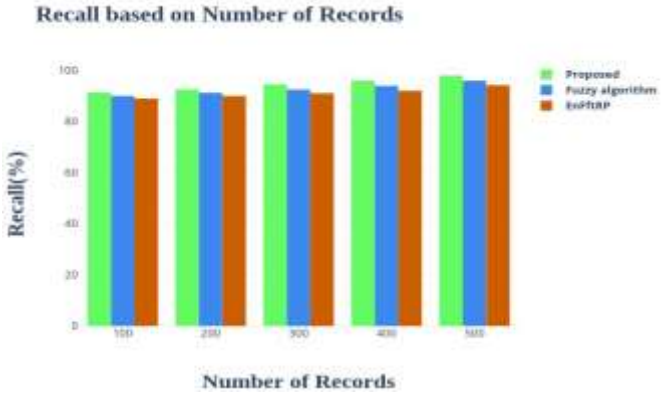

Fig. 8: Recall based on the Number of Records

5.2.4 F1-score

The model's accuracy in predicting students' education performance is measured by the F1 Score. As it illustrates the weighted harmonic mean of accuracy and recall, it indicates the balance between the precision and recall levels. The aim is for a better F1 score in students' education prediction, which is attained by having greater accuracy and recall values. For improved ranking prediction outcomes, the F1 score places a strong focus on increasing both accuracy and recall. It's derived as:

$$F1 - score = 2 * \left({}^{(\rho * \varepsilon)} \Big/ {}_{(\rho + \varepsilon)} \right) \quad (16)$$

In the equation (16) ρ refers to precision and ε refers to recall.

Table 6. Numerical Outcomes of F1-score (%)

No of Records	F1-score (%)		
	Proposed	Fuzzy algorithm	EnFftRP
100	95.5	94	92
200	96	95.2	93.5
300	96.4	95.8	94
400	97.2	96	94.7
500	98.4	97.1	95.5

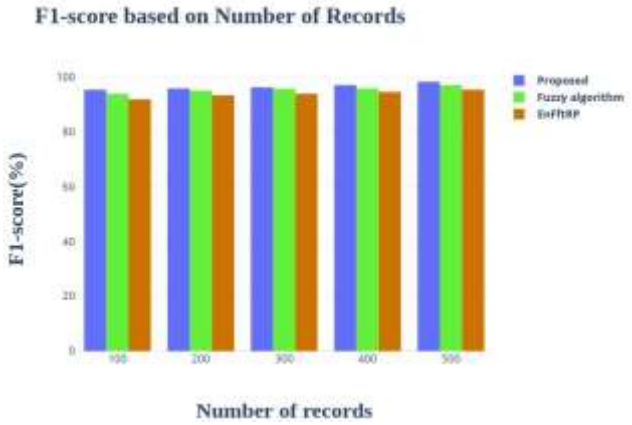

Fig. 9: F1-score based on Number of Records

Figure 9 and Table 6 indicate the outcomes of the F1-score. Comparing the proposed to the existing methods, such as 97.1% for the Fuzzy algorithm, 95.5% for EnFftRP, and 98.4% for the proposed. From this, it is clearly understood that the proposed has a higher F1-score. When the F1 score is higher for the prediction of students' academic success, the model's predictions show a better balance between recall and accuracy. By decreasing the number of false positives and false negatives, the model can accurately identify students who would do well.

5.2.5 Sensitivity

Sensitivity is an important evaluation criterion in machine learning and classification algorithms, which is the ratio of true positives among the total number of actual positives. This is especially important in problems where false negatives can be especially dangerous – for detecting frauds, diseases, or students needing additional attention. Sensitivity is calculated using the formula (Equation 17): Sensitivity is calculated using the formula (Equation 17):

$$Sensitivity = {T_rP_{ro}}/{(T_rP_{ro} + F_sNP_{ro})} \quad (17)$$

Table 7. Numerical Outcomes of Sensitivity (%)

No of Records	Sensitivity (%)		
	Proposed	Fuzzy algorithm	EnFftRP
100	95.2	93.5	92
200	95.7	94	93.2
300	96.8	95.1	94
400	97.2	95.9	94.9
500	98.4	97.3	95.6

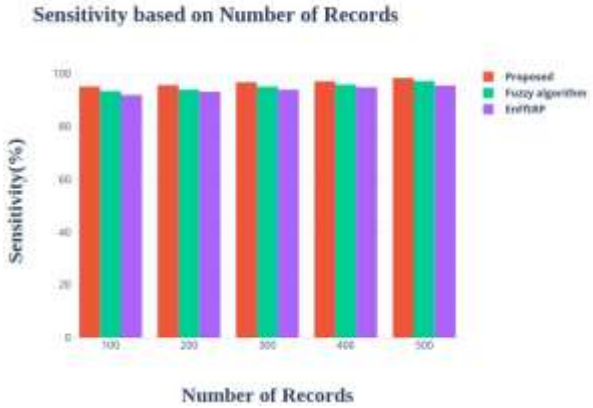

Fig. 10: Sensitivity based on Number of Records

Figure 10 shows Sensitivity results indicate that the Proposed Algorithm has the biggest sensitivity for all record numbers, from 95.2% at 100 records to 98.4% for 500 records. This shows that the proposed model is very efficient in maximizing the True Positive and minimizing the False Negative rates. The algorithm remains highly accurate even with greater records showing that it can scale well if more records are introduced. The Fuzzy Algorithm yields comparatively low results, slightly lower than the Proposed Algorithm, but with good sensitivity values ranging from 93.5% with 100 records to 97.3% for 500 records. However, it is always inferior to the proposed method by 1% to 2%, which could indicate a somewhat higher false negative rate of the algorithm. Concerning sensitivity, it is clear that EnFftRP is rated the lowest in all the datasets, with the sensitivity rates ranging from 92% in the first record to 95.6% in 500 records. Although the growth exceeds the others, its capacity to detect positive cases remains low. It is comparatively weaker in providing minimum false negatives and, therefore, may miss out on vulnerable cases. From the results, it can be said that for practical applications, the Proposed Algorithm is always the preferred choice in terms of sensitivity; at the same time, the Fuzzy Algorithm can be employed as an alternative when

the Proposed Algorithm is unavailable or when the model has to deal with certain fuzzy logic inputs.

5.2.6 Specificity

The degree of specificity indicates how well a prediction system selects ineffective individuals. It is the proportion of truly non-productive values among all the values that have no circumstances. Equation (18) is used to determine it.

$$Specificity = {}^{T_rNP_{ro}}/_{(F_sP_{ro} + T_rNP_{ro})} \quad (18)$$

Table 8. Numerical Outcomes of Specificity (%)

No of Records	Specificity (%)		
	Proposed	Fuzzy algorithm	EnFftRP
100	96.2	95	93.5
200	96.9	95.5	94
300	97.1	96.1	94.2
400	97.8	96.9	95
500	98.3	97.2	95.4

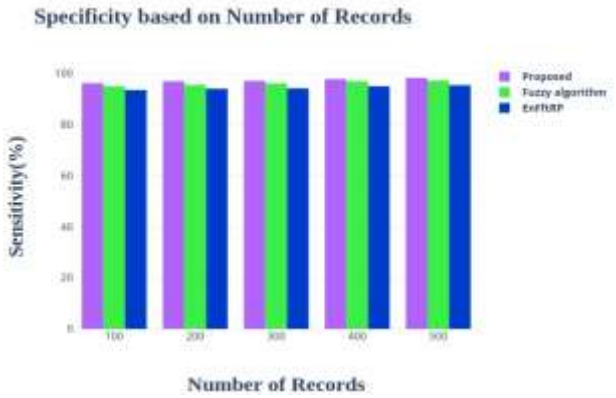

Fig. 11: Specificity based on the Number of Records

Figure 11 and Table 8 indicate the outcomes of specificity. Comparing the proposed to the existing methods, such as 97.2% for the Fuzzy algorithm, 95.4% for EnFftRP, and 98.3% for the proposed method have higher specificity than the existing techniques. When predicting students' educational achievement, a greater specificity metric means the model is more adept at recognizing real negatives that is, situations in which the model accurately predicts students who would struggle academically.

Table 9. F1-score based on Number of Records

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	85	80	75	77
Random Forest	90	85	83	84
Support Vector Machine	87	83	80	81
Logistic Regression	82	78	76	77
Neural Networks	92	88	86	87
Gradient Boosting	89	84	82	83

Fig. 12: Performance Metrics for Different Models

Table 9 and Figure 12 includes hypothetical results of applying different algorithms (e.g., Decision Tree, Random Forest, Support Vector Machine) and tracking key performance metrics such as Accuracy**,** Precision**,** Recall**,** and F1-Score**.** The analysis establishes that Neural Networks are the most accurate algorithms in identifying vulnerable students with an accuracy rate of 92%, followed by Random Forest at 90%, Support Vector Machine at 83% precision, Gradient Boosting at 84%, Decision Tree at 82% and 82% for Logistic Regression. These algorithms perform better than others in simpler problems; the second best is Gradient Boosting.

5.3 Research Summary

First of all, we read the required data from the necessary database. The last step will follow this: data pre-processing and exploratory data analysis. In this regard, the raw data is processed to learn more about it and its characteristics and detect trends and occurrences. Thirdly, data pre-processing may involve deleting missing values, removing outliers,

and normalizing our data values for use in the model. Handling Missing Values: Mean imputation should be used when one is forced to replace missing values with the mean of the variables in question. This method applies to those variables that are normally distributed. Addressing outliers: A logarithm of data is to be taken to limit the output impact of outliers and scale down the range of dynamics. Normalization: Normalization is one of the data pre-processing techniques used to scale data to make it to a certain standard for analytical or comparative purposes. Next, there is Feature extraction to choose the variable features such as student behavioral reports, the description of the course, and assessments from the documents; this improves the conversion of the document to the document-topic distribution through the use of GA-LDA. After that, interpretation and evaluation are done. Clustering the data has two main phases: A brief look at the two procedures involved: initialization and clustering. Initialization phase: In this phase, we define students' canopies and calculate T1 and T2 as distances where students belong to canopies. Clustering Phase: That is where students should be divided into and get the outcome from this teaching method.

We are detecting academic performance patterns through Gaussian Flow Optimizer Statistical Modelling. Categorizing the data— Enhance the prediction results of at-risk students using the Logistic Regression with K-Nearest Neighbour (LR-KNN) on the student's behavior. After that, we Conduct Early intervention and support, where we identify academic and non-academic challenges, implement IEPs, and modify them.

The latter requires continuous supervision of the work progress. The following performance indicators claim the efficiency of the suggested approach: Number of Records to Accuracy, Number of Records to Precision, Number of Records to Recall, Number of Records to F1-score, Number of Records to Sensitivity, Number of Records to specificity. This subsection also evaluates the suggested strategy and analyzes its performance. The above comparative results are presented graphically, as shown in Figure 6 and, Figure 11 and Table 3, Table 4, Table 5, Table 6, Table 7 and Table 8.

6 Conclusion

The ultimate goal is to enhance students' performance, decrease the number of academic dropouts, develop a support set of plans, and adapt these plans by continuous progress control. Moreover, it tackles the issues of overcomplicating over-complicating the relationships in a given predictive model, the complexity of interactions between variables, the inadequacy of feature space exploration, and data pre-processing intricacies. We help in Data Gathering, Exploration, Data cleaning, feature engineering, Insight & Analysis (grouping the data, categorizing the data), Counselling & Prompt Action, and Assistance. The simulation tool Python-3 established the factors influencing the correlation of visas with tourism. 11. 3, the model that is proposed is tested. The above-stated strategy is justifiable by weighing new techniques against the conventional methods with an accuracy of 97%. 9%, precision at 98. As for the stand-back tests, the recall is 2 %, and for the F1-score, the results are 98 %, respectively. 4%, sensitivity at 98. 4%, and specificity at 98. 3%. An evaluation of the performance of a strategy is done numerically, and it is seen that our approach outperforms the competition in every parameter. The size of the current dataset can be increased even more to improve the various prediction indicators of the subsequent works. In addition, applying the feature selection process can reduce the number of characteristics and determine the aspects that are significant to a student's performance.

References:

[1] Alsulami, A. A., AL-Ghamdi, A. S. A. M., & Ragab, M. (2023). Enhancement of E-Learning Student's Performance Based on Ensemble Techniques. *Electronics*, *12*(6), 1508. DOI: 10.3390/electronics12061508.

[2] Gil, P. D., da Cruz Martins, S., Moro, S., & Costa, J. M. (2021). A data-driven approach to predict first-year students' academic success in higher education institutions. *Education and Information Technologies*, *26*(2), 2165-2190. https://doi.org/10.1007/s10639-020-10346-6.

[3] Deem, R., Case, J. M., & Nokkala, T. (2022). Researching inequality in higher education: tracing changing conceptions and approaches over fifty years. *Higher Education*, *84*(6), 1245-1265. https://doi.org/10.1007/s10734-022-00922-9.

[4] Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational research*, *61*(3), 257-273. https://doi.org/10.1080/00131881.2019.1625716.

[5] Nikolaidis, P., Ismail, M., Shuib, L., Khan, S., & Dhiman, G. (2022). Predicting student attrition in higher education through the determinants of learning progress: A structural equation modelling approach. *Sustainability*, *14*(20), 13584. https://doi.org/10.3390/su142013584.

[6] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, *7*(11), 146. https://doi.org/10.24432/C5MC89.

[7] Singh, H. P., & Alhulail, H. N. (2022). Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach. *IEEE Access*, *10*, 6470-6482. DOI: 10.1109/ACCESS.2022.3141992.

[8] Amjad, S., Younas, M., Anwar, M., Shaheen, Q., Shiraz, M., & Gani, A. (2022). Data mining techniques to analyze the impact of social media on academic performance of high school students. *Wireless Communications and Mobile Computing*, *2022*, 1-11. https://doi.org/10.1155/2022/9299115.

[9] Makki, A. A., Sindi, H. F., Brdesee, H., Alsaggaf, W., Al-Hayani, A., & Al-Youbi, A. O. (2022). Goal programming and mathematical modelling for developing a capacity planning decision support system-based framework in higher education institutions. *Applied Sciences*, *12*(3), 1702. https://doi.org/10.3390/app12031702.

[10] Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, 19558-19571. DOI: 10.1109/ACCESS.2022.3151652.

[11] Musaddiq, M. H., Sarfraz, M. S., Shafi, N., Maqsood, R., Azam, A., & Ahmad, M. (2022). Predicting the impact of academic key factors and spatial behaviors on students' performance. *Applied Sciences*, *12*(19), 10112. https://doi.org/10.3390/app121910112.

[12] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*, 100066. https://doi.org/10.1016/j.caeai.2022.100066.

[13] Segura, M., Mello, J., & Hernández, A. (2022). Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?. *Mathematics*, *10*(18), 3359. https://doi.org/10.3390/math10183359.

[14] Sreenivasulu, M. D., Devi, J. S., Arulprakash, P., Venkataramana, S., & Kazi, K. S. (2022). Implementation of latest machine learning approaches for students grade prediction. *Int. J. Early Child*, *14*(3). DOI: 10.9756/INT-JECSE/V14I3.1141.

[15] Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J. (2023). Data mining and machine learning retention models in higher education. *Journal of College Student Retention: Research, Theory & Practice*, *25*(1), 51-75. DOI: 10.1177/1521025120964920.

[16] Pilotti, M. A., Nazeeruddin, E., Nazeeruddin, M., Daqqa, I., Abdelsalam, H., & Abdullah, M. (2022). Is initial performance in a course informative? Machine learning algorithms as aids for the early detection of at-risk students. *Electronics*, *11*(13), 2057. https://doi.org/10.3390/electronics11132057.

[17] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z.

[18] Albreiki, B., Habuza, T., Shuqfa, Z., Serhani, M. A., Zaki, N., & Harous, S. (2021). Customized rule-based model to identify at-risk students and propose rational remedial actions. *Big Data and Cognitive Computing*, *5*(4), 71. https://doi.org/10.3390/bdcc5040071.

[19] Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences*, *11*(7), 3130. https://doi.org/10.3390/app11073130.

[20] Nimy, E., Mosia, M., & Chibaya, C. (2023). Identifying At-Risk Students for Early Intervention—A Probabilistic Machine Learning Approach. *Applied Sciences*, *13*(6), 3869.

[21] Agarwal, N., & Tayal, D. K. (2022). FFT based ensembled model to predict ranks of higher educational institutions. *Multimedia Tools and Applications*, *81*(23), 34129-34162. https://doi.org/10.1007/s11042-022-13180-9.

[22] Yu, C. C., & Wu, Y. (2021). Early warning system for online stem learning—a slimmer approach using recurrent neural networks. *Sustainability*, *13*(22), 12461. https://doi.org/10.3390/su132212461.

[23] Sisa, I., Garcés, M. S., Crespo-Andrade, C., & Tobar, C. (2023, January). Improving Learning and Study Strategies in Undergraduate Medical Students: A Pre-Post Study. In *Healthcare* (Vol. 11, No. 3, p. 375). MDPI. https://doi.org/10.3390/healthcare11030375.

[24] Siddique, A., Jan, A., Majeed, F., Qahmash, A. I., Quadri, N. N., & Wahab, M. O. A. (2021). Predicting academic performance using an efficient model based on fusion of classifiers. *Applied Sciences*, *11*(24), 11845. https://doi.org/10.3390/app112411845.

[25] Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, *9*, 7519-7539. DOI: 10.1109/ACCESS.2021.3049446.

[26] Pek, R. Z., Özyer, S. T., Elhage, T., Özyer, T., & Alhajj, R. (2022). The role of machine learning in identifying students at-risk and minimizing failure. *IEEE Access*, *11*, 1224-1243. DOI: 10.1109/ACCESS.2022.3232984.

[27] Yang, X., & Ge, J. (2022). Predicting student learning effectiveness in higher education based on big data analysis. *Mobile Information Systems*, *2022*. DOI: 10.1155/2022/8409780.

[28] Casanova, J. R., Gomes, C. M. A., Bernardo, A. B., Núñez, J. C., & Almeida, L. S. (2021). Dimensionality and reliability of a screening instrument for students at-risk of dropping out from higher education. *Studies in Educational Evaluation*, *68*, 100957. DOI: 10.1016/j.stueduc.2020.100957.

[29] Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy*, *23*(4), 485. https://doi.org/10.3390/e23040485.

[30] Sáiz-Manzanares, M. C., Rodríguez-Díez, J. J., Díez-Pastor, J. F., Rodríguez-Arribas, S., Marticorena-Sánchez, R., & Ji, Y. P. (2021). Monitoring of student learning in learning management systems: An application of educational data mining techniques. *Applied Sciences*, *11*(6), 2677. https://doi.org/10.3390/app11062677.

[31] Liu, T., Wang, C., Chang, L., & Gu, T. (2022). Predicting High-Risk Students Using Learning Behavior. *Mathematics*, *10*(14), 2483. https://doi.org/10.3390/math10142483.

[32] Hussain, S., & Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of data science*, *10*(3), 637-655. https://doi.org/10.1007/s40745-021-00341-0.

[33] Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences*, *10*(11), 3894. https://doi.org/10.3390/app10113894.

[34] Shreem, S. S., Turabieh, H., Al Azwari, S., & Baothman, F. (2022). Enhanced binary genetic algorithm as a feature selection to predict student performance. *Soft Computing*, *26*(4), 1811-1823. https://doi.org/10.3390/app10113894.

[35] Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access*, *9*, 140731-140746.

[36] Queiroga, E. M., Batista Machado, M. F., Paragarino, V. R., Primo, T. T., & Cechinel, C. (2022). Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay. *Information*, *13*(9), 401. https://doi.org/10.3390/info13090401.

[37] Chen, S. (2022). Improved fuzzy algorithm for college Students' academic Early Warning. *Mathematical Problems in Engineering*, *2022*. DOI: 10.1155/2022/5764800.

[38] Mohd Talib, N. I., Abd Majid, N. A., & Sahran, S. (2023). Identification of Student Behavioral Patterns in Higher Education Using K-Means Clustering and Support Vector Machine. *Applied Sciences*, *13*(5), 3267. https://doi.org/10.3390/app13053267.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, KSA, under the grant no. G-691-135-37. The authors, therefore, acknowledges with thanks DSR technical and financial support.

Conflict of Interest

The authors have no conflicts of interest to declare.

APPENDIX

Table 1. Research Gap in Literature Survey

Reference	Objectives	Methods or Algorithms Used	Limitations
[16]	To examine the relationship between a student's total success after the semester and their early performance in the course, specifically their grades on the first assignment and exam.	KNN, LR, NB, RF, SVM	Not applicable generally to other academic subjects and course types since they are restricted to the sample of selected courses.
[17]	To predict academic performance based on previous achievement grades.	RF, NN, LR, SVM, NB and kNN	Not accurately representing the ranges in student performance variability
[18]	To identify students who are in danger as soon as possible so that the most significant and impactful characteristics in the students' data are taken into account while implementing the proper remedial procedures.	Rule-based model	Ignore other factors affecting student performance, and treatment effectiveness depends on each student's specific circumstances.
[19]	To highlight the limits of the educational data datasets that are now accessible, the significance of data comprehension, and the data collection phase	LR, DT, NB, SVM, RF	The enormous amount of data involved places a burden on computer resources and slows down processing.
[20]	To determine the characteristics that are used to identify at-risk students.	Extra Tree Classifier (ETC)	Scalability issues
[21]	The establishment of statistical methods that make it possible to detect student attrition early.	ANN, SVM, KNN, DT and LR.	Specialized knowledge and large computing resources are needed.
[22]	To determine whether or not students will pass the course and how early forecasts can be made without sacrificing sufficient accuracy.	Recurrent Neural Network (RNN)	It will not perform well when there is a lack of data.
[23]	To illustrate how undergraduate medical students from a Latin American medical school were able to enhance their learning and study skills via a designed interventional program.	Learning and Study Strategies Inventory (LASSI) test	It is not suitable for all medical school scenarios or student populations.
[24]	To identify the crucial elements affecting secondary school students' success	Multilayer Perceptron (MLP), J48 algorithm, PART Classifier	The accuracy of its conclusions is impacted by incorrect data.
[25]	To develop the early potential detection of students who may become dropouts.	RF	Large amounts of computing power are required for real-time model execution.
[26]	To enhance high school students' education and detect risky students before the end of the schooling cycle.	RF, K-NN, DT, SVM, NB, LR and AdaBoost	Feature engineering impacted the ability to comprehend and generalizability of the model.
[27]	To integrate the teaching process, learning activities, and assessment activities to promote student growth.	Indicator Analysis	Predictive algorithms are difficult to incorporate into existing protocols and educational systems.
[28]	The creation and approval of a method to evaluate first-year university dropout risk	Dimensionality test	It doesn't provide a thorough analysis of how effectively it indicates which students will ultimately drop out.
[29]	To develop ML algorithm-based models that collect appropriate information for forecasting different degrees of student retention,	DT, k-NN, LR, NB, RF, and SVM	Risk of overfitting.
[30]	To forecast students' academic development according to their study patterns.	k-means ++, fuzzy k-means, and Density-based spatial clustering of applications with noise (DBSCAN)	Restricted the evaluation of its efficacy across all academic courses and disciplines.
[31]	To forecast student success using the way that students learn.	Deep Neural Network (DNN), Simple Recurrent Network (SRN), Long Short-Term Memory Neural Network (LSTM), Gate Recurrent Unit Neural Network (GRU), and Convolutional Neural Network (CNN).	Requires higher computational resources.
[32]	To evaluate the level of education that is directly associated with the objectives of sustainability	Genetic algorithm, K-Fold Cross-Validation	Limited handling of time series data.
[33]	To apply DM and video learning analytics to forecast students' final performance for the semester.	NB, KNN, LR, ANN, SVM, RF, DT	Limited Interpretability.
[34]	To create an effective framework that can	Enhanced binary genetic algorithm, kNN,	Underfitting issues.

Reference	Objectives	Methods or Algorithms Used	Limitations
	forecast student grades using predetermined data linked to the experiment of the students.	DT, NB, SVM, and LDA	

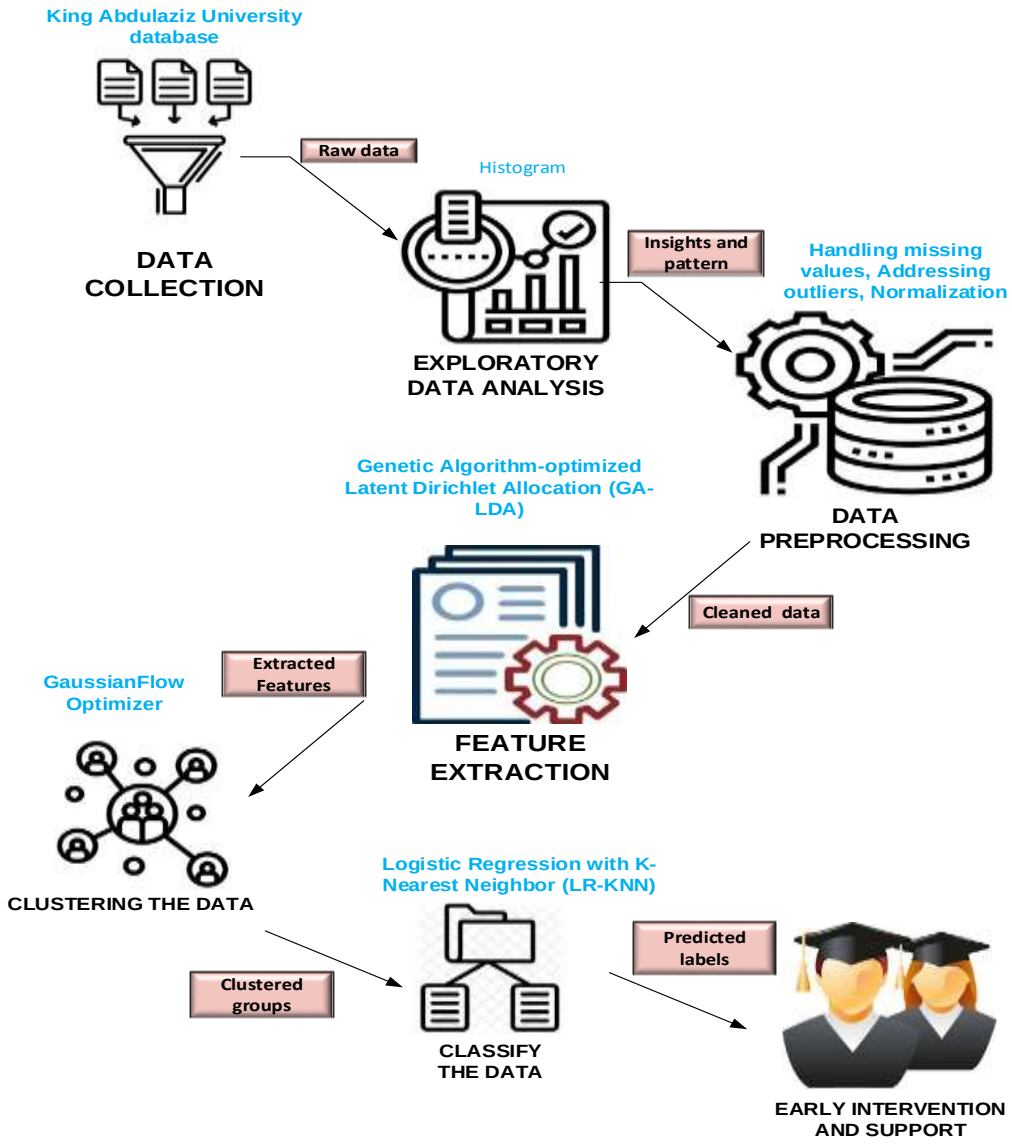

Fig. 1: Proposed architecture