# A WEB BASED PREDICTIVE MODEL FOR ACADEMIC PERFORMANCE OF THE STUDENTS

Thesis Submitted for the Award of the Degree of

**DOCTOR OF PHILOSOPHY**

**in**

**Computer Applications**

**By**

**Kajal Mahawar**

**Registration Number: 42100108**

**Supervised By**

**Dr. Punam Rattan (26558)**

**School of Computer Applications (Professor)**

**Lovely Professional University**

Transforming Education Transforming India

**LOVELY PROFESSIONAL UNIVERSITY, PUNJAB**
**2026**

# DECLARATION

I hereby declare that the work presented in this thesis, entitled "A Web Based Predictive Model for Academic Performance of the Students", submitted in fulfillment of the requirements for the degree of **Doctor of Philosophy (Ph.D.)**, is the outcome of original research carried out by me under the supervision of Dr. Punam Rattan, Professor, School of Computer Applications, Lovely Professional University, Punjab, India. In keeping with the general practice of reporting scientific observations, due acknowledgements have been made wherever the work described is based on the findings of other investigators. This work has not been submitted, either in part or in full, to any other university or institute for the award of any degree.

**Kajal Mahawar**

Registration No.: 42100108

School of Computer Applications
Lovely Professional University,
Punjab, India

# CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled "A Web Based Predictive Model for Academic Performance of the Students", submitted in fulfillment of the requirements for the award of the degree of **Doctor of Philosophy (Ph.D.)** in the School of Computer Applications, is a bonafide record of original research work carried out by Kajal Mahawar, Registration No.: 42100108, under my supervision. It is further certified that no part of this thesis has been submitted for the award of any other degree, diploma, or equivalent course.

**Dr. Punam Rattan (26558)**

(Professor)
School of Computer Applications
Lovely Professional University,
Punjab, India

# ABSTRACT

In the era of data-driven education, predicting student academic performance has become increasingly vital for timely interventions and informed decision-making. However, the complexity of student-related data, which includes academic, demographic, and behavioral features, presents challenges in identifying at-risk students effectively. The high dimensionality and diversity of such data make manual analysis inefficient, emphasizing the need for an optimized machine learning approach.

To resolve this issue, this study proposes a systematic and optimized framework for predicting whether students will score above or below 60%, a critical benchmark for campus selection. The research begins with the identification of 70 features, including state-of-the-art and novel features, collected through a well-structured questionnaire. These features were categorized into nine groups: demographic factors, educational factors, behavioral patterns, time-management skills, contextual influences, health factors, socioeconomic status, interpersonal factors, and family dynamics. A preliminary study was initially performed to validate the questionnaire and refine the data collection process. Based on the preliminary findings, the questionnaire was redesigned to focus on the most significant features. Subsequently, sample size estimation using Cochran's formula guided large-scale data collection from multiple institutions. The data preparation process, namely data cleaning, transformation, integration, reduction, and discretization, was employed to ensure quality and consistency in the dataset.

Feature selection was carried out using nine techniques across filter, wrapper, and embedded methods, namely Feature Importance, Ridge, Pearson Correlation, Variance Threshold, XGBoost, Random Forest, Recursive Feature Elimination, Chi-square, and Lasso. Among these, the Chi-square method provided the most accurate results, leading to the selection of 21 optimal features (excluding the target feature). To further enhance model performance, the redesigned questionnaire based on these selected features was used for refined data inputs, and SMOTE was applied to handle data imbalance. The final dataset was then partitioned into separate sets for rigorous performance assessment.

Multiple machine learning classifiers namely Support Vector Machine, Decision Tree, XGBoost, Naïve Bayes, Support Vector Machine (linear kernel), Logistic Regression, K-Nearest Neighbors, and Random Forest were trained and evaluated. The Decision Tree model

outperformed others in prediction accuracy and was further optimized using the Artificial Bee Colony (ABC) technique followed by Ant Colony Optimization (ACO), resulting in an improved and reliable predictive model.

The final optimized model was deployed through a Django-based web interface that supports both single and multiple student predictions. Users can enter individual student details or upload Excel files for batch processing. The system generates results indicating whether a student is likely to score above or below 60% and also provides post-prediction recommendations, which can be downloaded by the user in PDF format.

This study contributes:

- A custom dataset with performance-related features from multiple institutions.
- An optimized prediction model (ACO-DT) tailored for academic success classification.
- A real-time web-based system for academic performance forecasting and personalized guidance.

By combining advanced feature selection, machine learning, and bio-inspired optimization, this research presents a robust framework that supports educators in identifying and assisting at-risk students, thereby improving academic outcomes and refining campus recruitment strategies.

# ACKNOWLEDGMENT

It is not possible to prepare a project report without the assistance and encouragement of others, and this one is certainly no exception. At the very outset, I would like to express my heartfelt gratitude to all the individuals who supported me throughout this endeavor. I am especially and ineffably indebted to **Dr. Punam Rattan**, my guide, for her conscientious guidance, constant support, and encouragement throughout the course of this project. I benefited immensely from her logical analysis, insightful reviews, constructive feedback, and thoughtful comments.

Words are inadequate to express my gratitude to my father, **Mr. Harish Narayan Mahawar**, my mother, **Mrs. Anita Mahawar**, and my husband, **Mr. Swamesh Gupta**, for their prayers and unwavering support. Their constant encouragement and belief in me gave me the strength and confidence to complete my research. I am deeply thankful to my family for their cooperation, love, motivation, understanding, patience, and sincere advice.

I also express my heartfelt thanks to all the institutions that contributed to my research by facilitating data collection and validating my work. Their invaluable support and motivation have played a crucial role in the successful completion of this study.

Above all, I firmly believe that everything in this world happens according to the divine plan and will of the Almighty God. I thank him for guiding me and granting me the strength to carry out this research. I pray that he continues to shower his grace and blessings upon me and everyone around me.

Last but not least, my sincere gratitude goes to all my in-laws and friends who, directly or indirectly, helped me in completing this project report.

Kajal

**Kajal Mahawar**

# TABLE OF CONTENTS

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

| FIGURE NO. | LIST OF FIGURES | PAGE NO. |
|---|---|---|

# Appendices

## List of Appendices for All Journal Papers

## List of Appendices for All Conference Papers

## List of Appendices for the Institutions that Participated and Supported the Implementation and Validation of This Work

## List of Appendices for the Copy of Original and Revised Questionnaires

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ML | Machine Learning |
| AI | Artificial Intelligence |
| LR | Logistic Regression |
| SVM | Support Vector Machine |
| SVM-LK | SVM with Linear Kernal |
| RF | Random Forest |
| NB | Naïve Bayes |
| DT | Decision Tree |
| KNN | K-Nearest Neighbors |
| ACO | Ant Colony Optimization |
| SAPP | Student academic performance prediction |
| DL | Deep Learning |
| LMS | Learning Management System |
| RFE | Recursive Feature Elimination |
| GPA | Grade Point Average |
| XGB | XGBoost |
| RBF | Radial Basis Function |
| ABC | Artificial Bee Colony |
| SAP | Student Academic Performance |
| EDM | Educational Data Mining |
| DSS | Decision Support Systems |
| TEL | Technology-Enhanced Learning |
| NBC | Naive Bayes Classifiers |
| MATLAB | Matrix Laboratory |

# LIST OF ABBREVIATIONS

| | |
|---|---|
| WEKA | Waikato Environment for Knowledge Analysis |
| NN | Neural Networks |
| CART | Classification And Regression Tree |
| CHAID | Chi-Square Automatic Interaction Detection |
| ANN | Artificial Neural Networks |
| SNS | Amazon Simple Notification Service |
| DSSD | Data Structured Systems Development |
| KEEL | Knowledge Extraction based on Evolutionary Learning |
| SD-Map | Standard Definition Maps |
| RS | Recommendation Systems |
| UNSW-NB15 | UNSW Network Intrusion Detection |
| IDS | Intrusion Detection System |
| DO-IDS | Data Owner Intrusion Detection System |
| GA | Genetic Algorithm |
| VNS | Variable Neighborhood Search |
| GAVNS | Genetic Algorithm Variable Neighborhood Search |
| ARMCM | Augmented Reality-Based Multidimensional Concept Map |
| MCM | Multidimensional Concept Map |
| SPSS | Statistical Package for the Social Sciences |
| EPPI | Evidence for Policy and Practice Information |
| PRISMA | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| MBA | Master of Business Administration |
| CMS | Content Management System |

# LIST OF ABBREVIATIONS

| | |
|---|---|
| NMF | Non-negative Matrix Factorization |
| SVD | Singular Value Decomposition |
| PLS-DA | Partial Least Squares Discriminant Analysis |
| MOOCs | Massive Open Online Courses |
| LSTM | Long Short-Term Memory |
| RNN | Recurrent Neural Networks |
| CNN | Convolutional Neural Networks |
| GRU | Gated Recurrent Unit |
| MySQL | My Structured Query Language |
| NLP | Natural Language Processing |
| UK | United Kingdom |
| LOF | Local Outlier Factor |
| DBSCAN | Density-Based Spatial Clustering of Applications with Noise |
| ITB | Incremental Tree-Based |
| AVF | Adaptive Voting Framework |
| CBRW | Coupled Biased Random Walks |
| MCD | Minimum Covariance Determinant |
| RPCA | Robust Principal Component Analysis |
| OUVAS | Outlying Feature Value Set |
| RHAC | Random walk-based Hierarchical Value Couplings |
| AIEd | Artificial Intelligence in Education |
| DIEd | Deep Learning in Education |
| CNNs | Convolutional Neural Networks |

# LIST OF ABBREVIATIONS

| | |
|---|---|
| SC-GAN | Sequential Conditional Generative Adversarial Networks |
| SMOTE | Synthetic Minority Oversampling Technique |
| GAD | Generalized Anxiety Disorder |
| MDD | Major Depressive Disorder |
| EHRs | Electronic Health Records |
| AUC | Area Under the Curve |
| SHAP | Shapley Additive exPlanations |
| GMM | Gaussian Mixture Models |
| MLP | Multiple Linear Regression |
| PFA | Performance Factors Analysis |
| PNN | Probabilistic Neural Network |
| ADTree | Alternating Decision Tree |
| MAPE | Mean Absolute Percentage Error |
| ASD | Autism Spectrum Disorder |
| SVR | Support Vector Regression |
| BPNN | BP Neural Network |
| MSE | Mean Square Error |
| GBRT | Gradient Boosting Regression Tree |
| GS | Grid Search |
| SGD | Stochastic Gradient Descent |
| L-BFGS | Limited-memory Broyden-Fletcher-Goldfarb-Shanno |
| PSO | Particle Swarm Optimization |
| HRO | Hybrid Rice Optimization |

# LIST OF ABBREVIATIONS

| | |
|---|---|
| FIR | Finite Impulse Response |
| IIR | Infinite Impulse Response |
| BCO | Bee Colony Optimization |
| VAK | Visual, Auditory, and Kinesthetic |
| OULAD | Open University Learning Analytics Dataset |
| PHP | HyperText Preprocessor |
| TP | True Positive |
| TN | True Negative |
| FP | False Positive |
| FN | False Negative |
| ROC | Receiver Operating Characteristic |
| SAC | St. Aloysius' College |
| CSV | Comma Separated Values |
| FS | Feature Selection |
| CV | Cross-Validation |
| ABC-DT | Artificial Bee Colony-Decision Tree |
| ACO-DT | Ant Colony Optimization- Decision Tree |
| MLR | Multiple Linear Regression |
| LGBTQ+ | Lesbian, Gay, Bisexual, Transgender, Queer |

Chapter 1

# INTRODUCTION

A state’s development relies heavily on the academic success of its students. Strong performance cultivates skilled graduates who contribute to national progress. Predicting student performance is challenging due to diverse interpersonal, economic, educational, and social factors [1]. Industry requirements for placements demand high grades, posing employment risks for underperforming students [2]. Academic institutions increasingly invest in AI tools to estimate student success [3]. Various classification models, Logistic Regression (LR), Support Vector Machine (SVM), SVM with Linear Kernal (SVM-LK), Naïve Bayes (NB), XGBoost (XGB), Random Forest (RF), K-Nearest Neighbor (KNN), Decision Tree (DT), and others are used for performance prediction [4]. Additionally, optimization techniques like Ant Colony Optimization (ACO) enhance model accuracy and efficiency.

## 1.1 STUDENT ACADEMIC PERFORMANCE

Student academic performance prediction (SAPP) is a field of study that applies statistics, machine learning, artificial intelligence as well as deep learning to analyze educational data [5]. Governments and institutions worldwide are highly concerned about predicting a student's success. Anticipating grades enables educators to present students at-risk with appropriate warnings in advance, assisting them in averting issues and overcoming academic obstacles [6]. As a result, ML methods are being applied more and more to analyse student data and find hidden patterns associated with grade prediction.

### 1.1.1 Definition and Importance

**A. Definitions**

SAPP has been defined in various ways by researchers, each highlighting different aspects of the field:

According to [4], "Student academic performance prediction involves the application of machine learning and statistical techniques to forecast students' future academic outcomes based on historical data and identified influencing factors."

According to [7], "The prediction of student academic performance leverages artificial intelligence methods to analyze patterns in student data, providing insights into potential academic success and areas needing improvement."

According to [8], "Academic performance prediction is the process of using data mining and machine learning algorithms to predict students' grades and performance in their courses, enabling timely interventions and personalized support."

According to [9], "Student academic performance prediction utilizes advanced computational models to evaluate various predictors of academic success, aiming to provide actionable insights for educators and policymakers."

According to [10], "The study of predicting student academic performance focuses on employing deep learning techniques to identify and forecast students' academic trajectories, thereby assisting in the development of targeted educational strategies."

**B. Importance**

Academic performance prediction of students is important for higher education and beyond. It makes it possible to identify at-risk students early and to intervene quickly to stop learning difficulties and promote achievement [11]. Through the use of ML algorithms, hidden patterns in student data are revealed, enabling customised support based on each student's unique learning requirements. Precise forecasts enable educators and decision-makers to make well-informed choices, thereby optimising instructional methodologies and results. Refinement of curriculum design, instructional strategies, and resource allocation are made possible by an understanding of the elements that influence academic achievement [12]. This improves the quality of education as a whole. In the end, increased academic achievement makes students more marketable in the job market and successfully addresses issues of underemployment and unemployment [13].

### 1.1.2 Factors Influencing Academic Performance

Several factors influence student academic success. Understanding such factors is critical for making informed targeted actions and policy decisions. By examining crucial factors, educators and policymakers can devise methods to improve learning outcomes and close performance gaps [14]. These factors, such as demographic factors, educational background, behavior patterns, time-management skills, contextual influences, health factors, socioeconomic status, and family dynamics, collectively influence students' academic achievements, well-being, and future prospects [15]. The following are the factors:

### A. Demographic factors

To identify demographic factors, the features of individuals or groups within a population are considered. In the context of predicting a student's academic performance, demographic considerations typically encompass features like age, gender, religion, category, and so forth. These features can affect students' academic experiences and results as well as offer insights into the various backgrounds and learning situations in which they study [15], [16], [17], [18], [19], [20], [21], [22], [23].

### B. Educational factors

The learning environment and educational experiences that can affect a student's academic achievement are referred to as educational factors. These features range from the standard of instruction to the availability of resources and facilities at the college, involvement in extracurricular activities, surroundings, educational background, preferred method of learning, etc. [15], [17], [18], [20], [21], [24], [25], [26], [27], [28].

### C. Behavioral factors

Students' activities, attitudes, and habits that affect their academic achievement are referred to as behavioral factors. The aforementioned factors include features like motivation, emotional stability, involvement in extracurricular activities outside of the classroom, and so forth. Behavior has an enormous effect on how effectively the student learns, understands the subject, and does academically [15], [16], [18], [20], [21], [23], [24], [29].

### D. Family factors

Family factors include a variety of aspects of a student's home environment that may influence their overall well-being and educational performance. This factor includes, the type of family, the surroundings, the parents' marital status, and the involvement and support of the parents [15], [17], [19], [23], [24], [28].

### E. Socioeconomic factors

The social and economic circumstances that affect a student’s or family's prospects and well-being are referred to as socioeconomic factors. These factors include features such as, employment status, part-time work, income level, and scholarships. Socioeconomic factors have a big impact on how students perform academically and how their educational experiences and results are shaped [15], [16], [18], [20], [21].

### F. Contextual factors

Contextual factors encompass environmental influences and issues on student academic performance, including area student live, language barrier, gender inequality, etc. In order to foster equitable learning settings and advance educational fairness, it is imperative that these features be recognized and addressed [15], [17], [22], [26], [27], [29], [30].

### G. Interpersonal factors

The dynamics and relationships that students have with their classmates and the ways in which these ties can affect academic performance are referred to as interpersonal factors. Positive relationships can boost academic engagement and accomplishment by offering social support, motivation, and chances for collaborative learning. On the other hand, unfavorable peer interactions that result in social isolation, disagreements, or peer pressure can distract from academic attention and lower performance [15], [16], [18], [30].

### H. Time-management factors

The capacity to efficiently set aside time for social media, library, lab, and online gaming activities while simultaneously juggling additional responsibilities like job, college, or personal commitments is referred to as time management factors. Games and social networking sites can be major time wasters and sources of distraction if they are not used appropriately. Overuse of social media and video games might cause people to put off crucial activities, become less productive, and procrastinate [15].

### I. Health factors

Health factors encompass physical and mental well-being, including trauma, mental health, chronic illnesses, etc. To ensure the success of students, these factors must be given attention and support because they have a substantial impact on students' academic achievement [15], [17], [18], [24].

### 1.1.3 Challenges in Measuring Academic Performance

Measuring and comparing student academic performance is challenging due to multiple factors [31]. Standardized tests often fail to capture skills like independent thinking and critical analysis, favoring rote memorization instead [32]. Students excelling in experimental or practical learning methods may underperform in such tests. Educational quality varies across schools, with differences in funding, extracurricular opportunities, and access to technology significantly affecting performance [33]. Additionally, personal circumstances, mental health, and family support further influence outcomes [34]. Traditional assessments rarely account for these environmental or social factors [5]. Cultural and linguistic diversity among students can bias standardized evaluations, producing data that may not accurately reflect true abilities [35]. Also, teaching often becomes test-focused, prioritizing high scores over holistic education.

In conclusion, challenges in assessing academic performance include standardized testing limitations, educational inequalities, external influences, cultural biases, and curriculum constraints. Addressing these requires careful data collection and the use of feature selection techniques, discussed in the next section.

## 1.2 FEATURE SELECTION TECHNIQUES

In ML, feature selection involves identifying a subset of the most relevant features from a high-dimensional dataset in order to improve model performance and computational efficiency. There are two types of this process: **supervised** and **unsupervised**.

When using **supervised feature-selection**, the selection procedure makes use of the output feature (class label). The aim is to identify which group of features best predicts the target feature [36]. Additionally, **unsupervised feature-selection** techniques do not depend on the output feature [37].

The most informative features are retained while extraneous or redundant ones are removed through this process. The goal of this process is to reduce the number of features without losing meaningful information, thereby improving model interpretability, accelerating model execution, and enhancing model prediction accuracy. Common approaches include **filter**, **wrapper**, and **embedding** methods [38].

Figure 1.1 depicts the feature selection methods as given below:

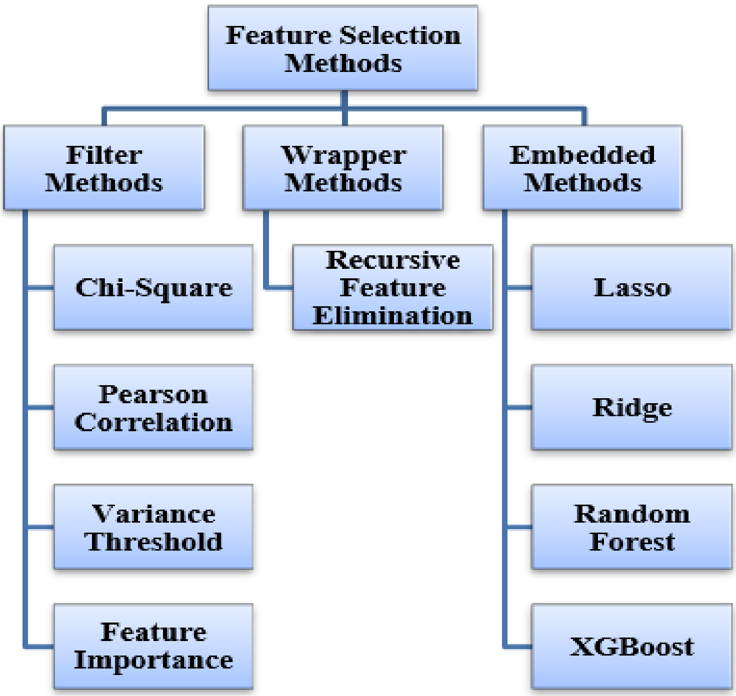


**Figure 1.1. Feature Selection Methods**

### 1.2.1 Filter method

The filter method for feature selection entails assessing each feature separately from the selected model and choosing those that satisfy predetermined standards according to their statistical characteristics or connection to the target variable [22]. Mathematically, it can be represented as,

$$\{Score\}(X\_i) = \{Measure\}(X\_i, Y) \tag{1.1}$$

where, ({Score}(X_i)) represents the score or relevance of feature (X_i) and

({Measure}(X_i, Y)) denotes the statistical measure of the relationship between feature (X_i) and the target variable (Y), such as correlation coefficient, mutual information, chi-square statistic, or others.

Figure 1.2 depicts the filter-feature selection methods as given below:

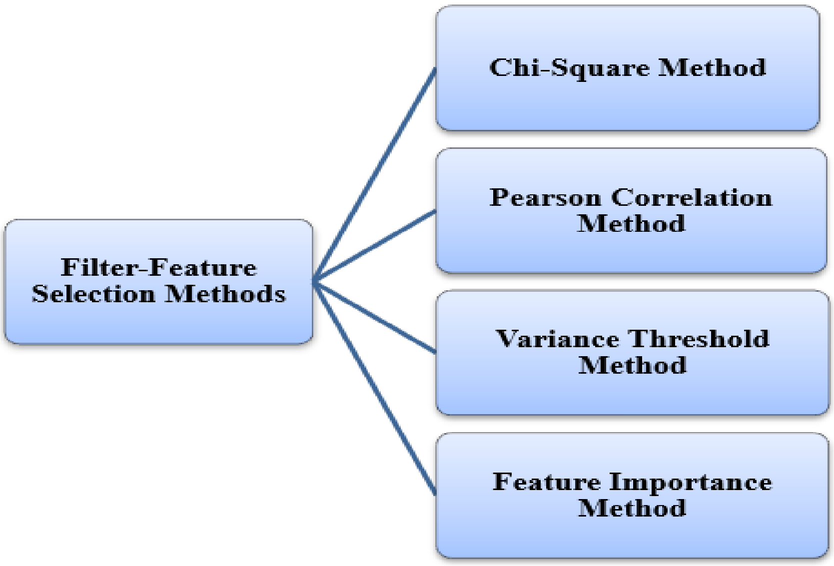


**Figure 1.2. Filter-Feature Selection Methods**

### A. Chi-Square method

The **Chi-Square method** ($x^2$) is a technique used to measure the affinity between non-numerical features. The test calculates the chi-square value, which measures how the observed frequencies in each category differ from the expected frequencies if there was no association. Features with higher chi-square statistics are considered more relevant for predicting the target feature [39].

### B. Pearson Correlation method

The **Pearson Correlation method** calculates the straight-line connection between two continuous features, expressed as a coefficient between -1 and 1. A coefficient of 1 signifies a full positive dependency, -1 signifies a full negative dependency, and 0 indicates no linear dependency. In feature selection, this method helps identify and eliminate features that are strongly correlated, reducing predictor collinearity [40].

### C. Variance Threshold method

The **Variance Threshold method** is a simple approach to feature selection that discards features with variance below a specified threshold. Variance quantifies how much a feature varies among different samples. Features with low variance are often less informative for differentiating between classes and can be removed. This technique is especially effective for eliminating features that provide minimal information [41].

### D. Feature Importance Method

The **Feature Importance method** involves assigning values to input features that contribute to predicting the target variable. Techniques used to determine feature importance include decision trees, random forests, and gradient boosting models [39].

## 1.2.2 Wrapper method

The wrapper technique uses model training and testing to evaluate multiple feature subsets. The subset that maximizes model performance is selected at the end of the process [5]. Mathematically, this can be summarized as Eq. (1.2),

$$\{Score\}(X\_\{S\}) = \{Performance\}(M(X\_\{S\}, Y)) \tag{1.2}$$

where, {Score}(X_{S}) represents the score or evaluation of a subset of features (X_{S}), (M(X_{S}, Y)) denotes a predictive model trained with the subset of variables (X_{S}) and the target feature Y, and ({Performance}(M(X_{S}, Y))) refers to the performance of the ML model. Figure 1.3 depicts the wrapper-feature selection method as given below:

### A. Recursive Feature Elimination method

**Recursive Feature Elimination method** is used for feature selection that progressively narrows down the feature set by iteratively eliminating the least significant features. It works by fitting a model, removing the least important features, and then refitting the model. This cycle continues until the desired number of features is obtained [39], [41].

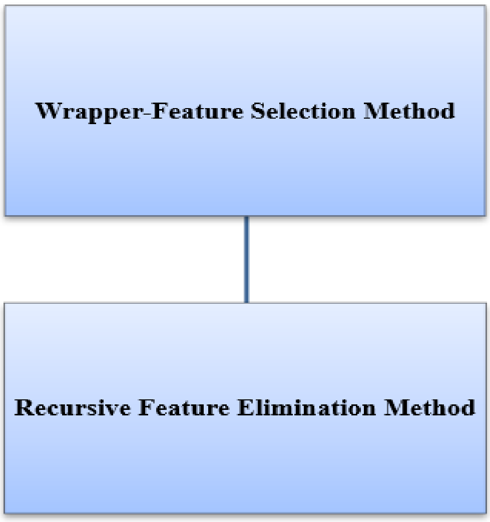


**Figure 1.3. Wrapper-Feature Selection Method**

### 1.2.3 Embedded method

Feature selection is incorporated into the model training process itself using embedded method (E), usually by penalising the coefficients of irrelevant features [5]. It can be expressed mathematically as Eq. (1.3),

$$E = \min\beta \text{Loss } (X, Y, \beta) + \lambda \cdot \text{Penalty}(\beta) \tag{1.3}$$

where, min$\beta$ represents the minimization of the loss function with respect to the model parameters $\beta$, Loss (X, Y, $\beta$) denotes the loss or error function, measuring the discrepancy between the model predictions and the true values Y given the features X and parameters $\beta$, $\lambda$ is a parameter controlling the penalty term, and Penalty($\beta$) represents the penalty term applied to the model parameters $\beta$, typically based on their magnitude. Common penalty terms include L1 (Lasso) or L2 (Ridge) penalties. Examples of this method include Elastic Net, Lasso regression, decision tree pruning, and Ridge regression. Figure 1.4 depicts the embedded-feature selection methods as given below:

#### A. Lasso method

A distinctive aspect of the Lasso method is its ability to shrink certain coefficients to zero, thereby automatically excluding irrelevant features [40]. This technique aims to minimize the residual sum of squares while imposing a constraint that the total absolute value of the coefficients remains below a specified limit.

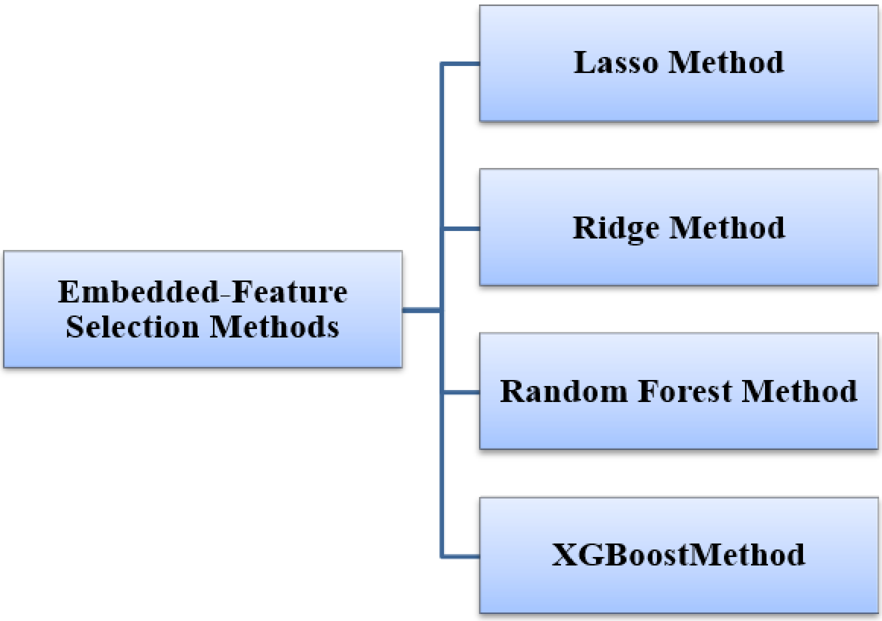


**Figure 1.4. Embedded-Feature Selection Methods**

### B. Ridge method

Unlike Lasso, Ridge regression doesn’t shrink coefficients to zero, so it doesn’t perform feature selection but instead reduces the size of coefficients to handle multicollinearity among features [42]. The objective is to minimize the residual sum of squares while adding a penalty proportional to the sum of the squared coefficients, thereby stabilizing the estimates.

### C. Random Forest method

Random Forest proves its worth through feature selection because it uses an ensemble learning approach that excels in the feature selection process. The model achieves better prediction stability and accuracy through its combination of many trees and merging their prediction results [43].

### D. XGBoost method

XGBoost functions as a reliable gradient boosting-based algorithm to execute both feature selection and model training procedures [44]. The necessity of features for a model depends on its impact in reducing loss function values across the ensemble of trees.

A combination of selected features and ML algorithms produces results according to the above-described methods. Through accurate predictions, the method determines student performance levels. The next section details an extensive description of multiple machine learning methods that power forecasting operations.

## 1.3 MACHINE LEARNING TECHNIQUES

Students' academic performance can be accurately predicted using Machine Learning (ML) methods. Models utilize various data sources, including attendance, grades, socioeconomic status, and behavioral patterns, to forecast achievement or failure [45]. The main goal is to identify at-risk students early, enabling timely interventions and support. The choice of ML techniques depends on the prediction task, whether it involves classifying students into performance groups (e.g., pass/fail, high/low achievers) or predicting continuous outcomes like GPA or test scores [46].

ML can be broadly classified into several categories depend on the training-learning approach and the nature of the tasks it is designed to perform. Figure 1.5 represents the categories of ML:

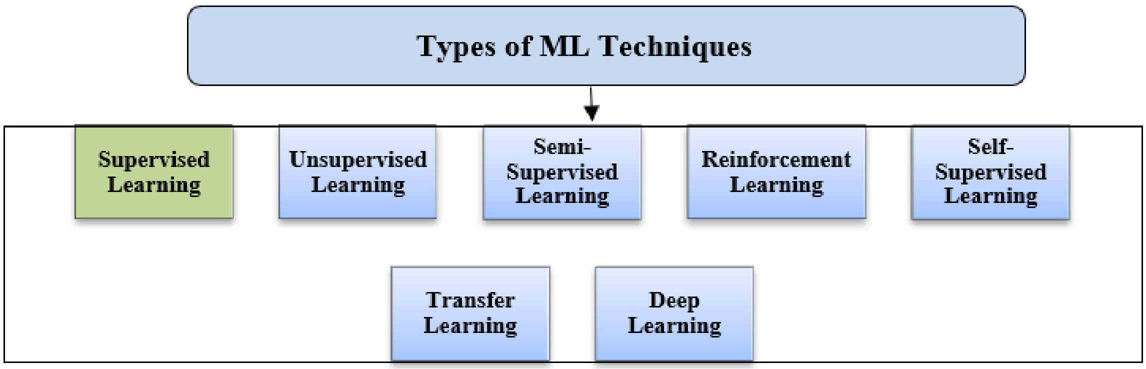


**Figure 1.5. Types of ML Techniques**

### 1.3.1 Supervised Learning

In supervised learning, models learn from labeled datasets, with each input mapped to a predefined output [3]. This method is particularly useful for predictive tasks, such as forecasting student academic performance. In this context, several ML techniques, such as Logistic Regression (LR), Decision Tree (DT), XGBoost (XGB), Naive Bayes

(NB), Support Vector Machine (SVM) and its Kernel variant, K-Nearest Neighbors (KNN), and Random Forest (RF), are employed. By learning from historical data, where student features are linked to academic outcomes, these supervised learning models can accurately forecast future assessment [31].

Figure 1.5.1 represents the supervised ML techniques:

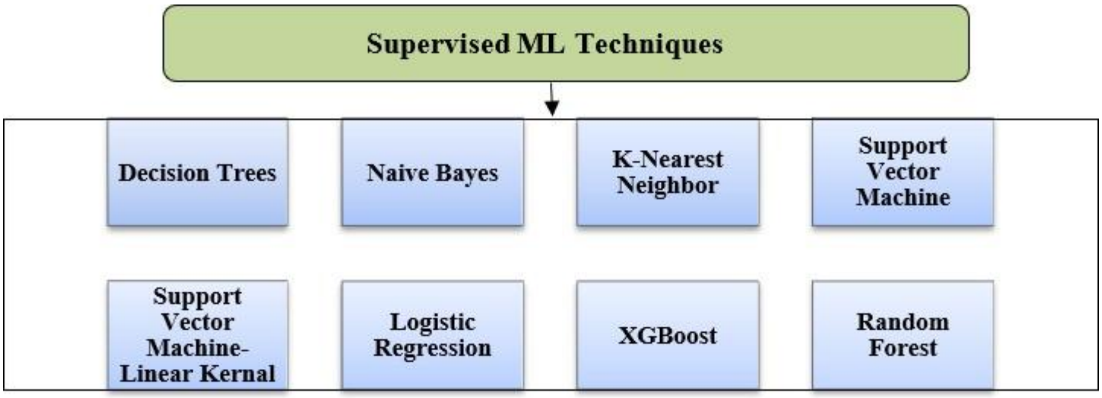


**Figure 1.5.1. Supervised ML Techniques**

Since many steps are common across different supervised learning models, a single ML workflow pipeline is demonstrated in Figure 1.5.2, illustrating the sequential process from data collection to deployment.

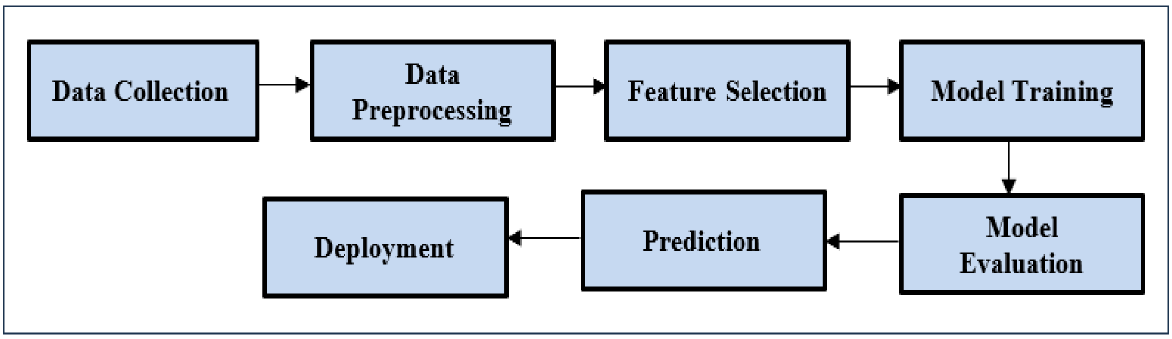


**Figure 1.5.2. ML Workflow Pipeline**

The process begins with data collection, followed by data preprocessing, which involves cleaning, transforming, and preparing the data for modeling. Next, feature selection identifies the most relevant attributes to improve model performance, and model training applies machine learning classifiers. The trained models are then assessed during model evaluation using appropriate metrics, followed by performance prediction. Finally, the optimized models are implemented in a web-based deployment to provide real-time predictions and recommendations for students at risk. Following this unified workflow, the study employs individual supervised learning models, such as:

### A. Decision Tree

Decision Trees (DT) build a model by repeatedly dividing data depend on feature values. Each split creates a node, and branches represent decision paths leading to leaves that indicate predicted outcomes [47]. The decision-making process and the value of features may be better understood with the aid of the tree structure. Pruning trees helps keep them from becoming too complex and overfitted. The model predicts outcomes by following decision rules. The purpose of evaluation is to measure the model's efficacy using several criteria [48].

### B. Naïve Bayes

Assuming that variables are independent when class labels are taken into consideration, Naïve Bayes (NB) uses Bayes' theorem as a probabilistic classification technique [26]. The procedure begins with preprocessed data that has features chosen using techniques like Chi-Square. Then, prior probabilities of classes and likelihoods of features are estimated. After that, the model is trained using Bayes' theorem to determine the probabilities of each class, and then predictions are made using the posterior probability with the greatest value. Finally, the accuracy and consistency of the model are assessed [38].

### C. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) classification algorithm finds the examples that are geographically nearest to a given data point using metrics such the Manhattan or Euclidean distance, and then assigns the data point to the class that has the most neighbors [49]. After pre-processing phase, distance between data points are then calculated, and the nearest neighbors are identified. Using the labels or values of these neighbors, the data point is either classed or its value is projected. Accuracy in classification and mean squared error in regression are some of the measures used to assess performance.

### D. Support Vector Machine

Support Vector Machine (SVM) is a ML technique that builds a feature space hyperplane. The main goal is to identify the hyperplane that increases the margin between the nearest points of different classes [50]. Preprocessed data is used to initiate the procedure. The next step is to create a hyperplane that will divide the classes. If the data cannot be separated linearly, then kernel functions will be used to transform it into higher dimensions. When training the model, the goal is to maximize the margin between the classes. The model's accuracy is assessed by looking at the data points' positions on the hyperplane, which allows predictions to be made [51].

### E. Support Vector Machine-Linear Kernal

Support Vector Machine-Linear Kernel (SVM-LK) stands as an SVM variant which uses kernel functions to solve non-linear data separations. Through dimensional projection SVM-LK creates intricate decision boundaries which surpass the ability of linear SVM models to do so [52]. The method provides exceptional results when modeling intricate relationships that occur between features and class labels. After data preparation processes users must select the suitable kernel function (RBF or polynomial etc.) to address the non-linear relationship in the data. Model optimization happens by determining a non-linear decision frontier generated in the transformed space. Performance evaluation happens after predictions through the use of the non-linear decision boundary [5], [53].

### F. Logistic Regression

The probability-based binary classification task requires the machine learning technique Logistic Regression (LR) for its operation [53]. Users gain better insights into class membership probabilities and variable influence on results through the implementation of this method. Chi-Square analysis of processed data selects features as part of the procedure. The next operation takes place through logistic function modeling of probabilities. The computation of model coefficients makes use of optimization methods when training the model [54]. Predictions are obtained through class probability calculations that use threshold values for class assignment in decision making. Finally, performance is evaluated using different metrics.

### G. XGBoost

XGBoost (XGB) is a hybrid learning technique that uses gradient boosting to maximize the model’s accuracy by reducing a loss function and incorporates regularization to prevent overfitting. XGB is renowned for its high performance and scalability [44].

The process begins with preprocessed data and feature selection. Model initialization involves setting up base models, such as decision trees. The boosting process then iteratively adds models to correct errors from previous ones. The model is fitted by optimizing features and applying regularization to avoid overfitting [55]. Predictions are aggregated from all models to produce the final output, and performance is evaluated.

**H. Random Forest**

To increase the accuracy and resilience of predictions, Random Forest (RF) uses a hybrid ML approach that combines many decision trees. The final forecast is generated by combining the results of all the trees, which are trained using a randomly selected sample of characteristics and data [43]. Effectively handling a broad variety of data types, this strategy helps prevent overfitting. Various approaches are used to choose features from preprocessed data at the beginning of the process. The next step is to build several decision trees using bootstrapped data samples. Each tree is given a random selection of characteristics [40]. The final output is the result of averaging or majority voting all the trees' predictions. At last, the model's performance is evaluated.

## 1.4 OPTIMIZATION TECHNIQUES

In ML, optimization techniques are pivotal for enhancing and refining the efficacy of predictive models. These techniques aim to fine-tune model parameters, minimize errors, and maximize predictive accuracy. Typical algorithms, such as gradient descent, and evolutionary algorithms like genetic algorithms and particle swarm optimization are commonly employed in different ML applications [56]. By incorporating optimization techniques into ML workflows, researchers can enhance model efficiency, and ultimately enhance predictive performance across various applications. Notably, this research focuses on the following two prominent optimization techniques [15].

**A. Artificial Bee Colony**

The ABC algorithm aims for the best answers by imitating how honey bees' forage [57]. In ABC, potential solutions, bees, search the search space, assess their fitness using a predetermined objective function, and communicate with one another via a waggle dance-like mechanism. Over iterations, the algorithm converges towards the optimal solution by iteratively updating candidate solutions based on the best-performing solutions discovered so far [58]. Figure 1.6 depicts the three main components of the algorithm.

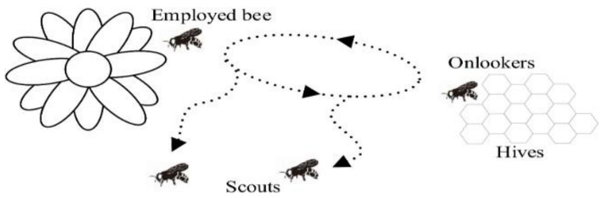


**Figure 1.6. The behavior of the bees**

### 1.4.1 Employed Bees

The employed bees are responsible for navigating the search space by utilizing current solutions while also uncovering new ones through focused, localized searches [59].

### 1.4.2 Onlooker Bees

These bees choose solutions using the data gathered from the employed bees and conduct localized searches to enhance these solutions [59].

### 1.4.3 Scout Bees

Scout bees are tasked with the role of randomly exploring new solutions, particularly when both employed and onlooker bees have depleted their search efforts [59].

**B. Ant Colony Optimization**

The ACO technique draws stimulus from the foraging behavior of ants and has been applied to diverse optimization problems [60]. While foraging for food, ants leave behind little amounts of a chemical known as pheromones on the ground. This pheromone helps ants retain their memories [61]. The actual ants' behavior is depicted in Figure 1.7. The actual ants in Figure 1.7. (A) travel a straight path from their colony to the food source. When an obstruction arises on the path in Figure 1.7. (B), ants have an equal chance of choosing to proceed left or right because there is no pheromone trail. At this point, some ants will select the top path, while others will select the bottom path. Pheromone is deposited faster on the shorter path in Figure 1.7. (C). After a while, as seen in Figure 1.7. (D), every ant has selected the shorter route [62].

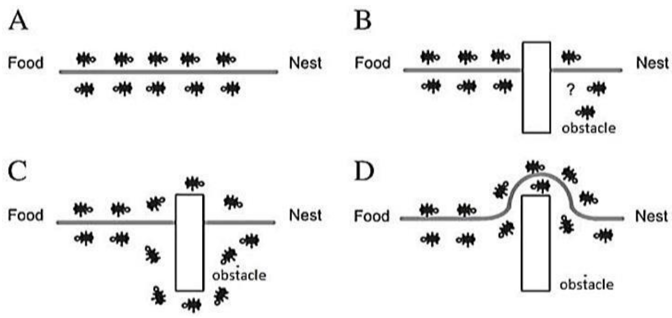


**Figure 1.7. The behavior of the actual ants'**

Both ABC and ACO algorithms are part of the metaheuristic optimization methods, renowned for their efficiency in navigating intricate solution spaces and identifying near-optimal outcomes [59].

**Novelty of the Study:** Unlike previous studies that primarily focused on final grades or dropout prediction, this research develops a web-based predictive model integrating multiple factors. By optimizing the model using Ant Colony Optimization, it accurately forecasts whether students achieve 60% or more. This approach provides timely, personalized recommendations for at-risk students and supports both on-campus and distance learning environments. Moreover, the model enables educators to identify students who need intervention early, facilitating targeted support and improving overall academic outcomes. Its deployment as a web-based system ensures accessibility, ease of use, and real-time performance monitoring.

## 1.5 DATA SOURCES AND COLLECTION

In this study, data were collected from 1,369 second-year undergraduate students in Jabalpur, Madhya Pradesh. A specifically designed questionnaire, distributed via Google Forms, served as the primary data collection tool. Prior to the main study, a pilot study with 142 students was conducted to test the questionnaire's reliability. The questionnaire was also validated by subject matter experts and English language experts via email to ensure clarity and relevance. From a total population of 3,186 students, a sample size of 1,369 was determined using the Cochran formula for finite populations [38].

**A. Feature Selection**

The dataset contains 70 features representing various factors influencing students' academic success. Optimal features were selected using multiple feature selection techniques, identifying the most relevant traits for predicting academic performance [15].

**B. Data Collection Process**

The questionnaire covered a wide range of factors, including demographic details, past academic performance, study patterns, and socio-economic background. Collected data were cleaned and organized to maintain accuracy and consistency. The primary outcome variable in this study is students' grades from the previous academic year, used to forecast academic success.

**C. Dataset Utilization**

The dataset was divided into three subsets: training, validation, and testing. The training set was used to build the predictive models, the validation set to fine-tune them, and the test set to evaluate the final model's performance. This structured approach enables the application of machine learning techniques to accurately predict student academic outcomes based on the selected features.

## 1.6 OVERVIEW OF THE PROPOSED SYSTEM

The overall workflow of the proposed student academic performance prediction system is illustrated in Figure 1.8.

This overview provides a clear context for the methodology described in later chapters. The figure illustrates the overall process, which begins with the collection of student data through questionnaires, followed by feature selection to identify the most relevant attributes for accurate predictions. The selected features are then used to train and optimize machine learning models using Artificial Bee Colony (ABC) and Ant Colony Optimization (ACO) techniques. Finally, the optimized model is deployed through a web-based interface to predict student academic performance ($\geq 60\%$ / $<60\%$) and provide recommendations for students who may be at risk.

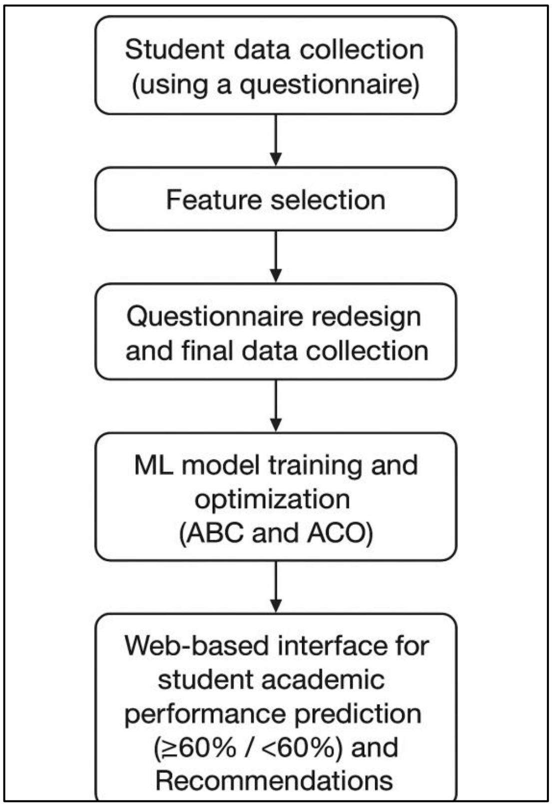


**Figure 1.8. Overview of the proposed student academic performance prediction system**

## 1.7 PROBLEM IDENTIFICATION

Machine Learning is an essential subset of AI that has significantly advanced predictive analytics across various domains, including education. In higher education, ML techniques have been extensively applied to predict student academic performance. However, most existing studies primarily focus on broad performance metrics without addressing the specific requirements of campus recruitment processes.

A critical benchmark for campus-based employment opportunities is whether a student secures 60% or more in academic performance, as many companies set this threshold as a minimum eligibility criterion for placements. Despite its importance, limited research has developed ML-based predictive models that explicitly assess students based on this requirement. Existing models often lack customized classification for employability, fail to incorporate key interpersonal, time-management, academic, health and family factors, or do not leverage advanced optimization techniques for improved accuracy. Additionally, no web-based predictive interface has been developed yet to identify at-risk students and provide recommendations based on these key factors.

To fill this void in existing research, this research focuses on:

- Identify the most influential predictive features impacting student performance, including interpersonal, time-management, academic, health and family factors.

- Develop a high-accuracy ML model that classifies students based on their eligibility for campus selection.

- Optimize the model's predictive performance using advanced metaheuristic techniques such as ABC and ACO algorithms.

- Design a web-based predictive interface to detect at-risk students and offer recommendations based on key influencing features.

By addressing these challenges, this research provides an effective decision-support tool for academic institutions and recruiters to detect at-risk students, enhance academic interventions, and streamline campus placement strategies.

## 1.8 AIM OF THE RESEARCH

In the realm of higher education, optimizing campus recruiting operations relies heavily on accurate predictions of students' academic success. The main purpose of this research is to establish an effective student classification system through ML prediction, which determines campus selection eligibility by estimating achievement levels above 60%. The major objective of this study involves enhancing academic prediction accuracy through identification of critical forecasting factors followed by advanced optimization techniques and machine learning strategies for prediction execution. The ultimate goal is to create a web-based interface for this predictive model, providing institutions with a reliable tool to identify students who meet the necessary academic standards for recruitment. Ultimately, a comparison of the model against modern standards will verify its accuracy for campus recruitment.

## 1.9 OBJECTIVES

- To identify the most influential predictive variables affecting students' academic performance in higher education.
- To formulate machine learning predictive model of selected parameters.
- To develop a web based predictive model interface.
- To test and validate proposed model with state-of-the-art models.

## 1.10 SCOPE OF THE STUDY

Researchers applied machine learning to automate the prediction process to build a forecasting model for student academic success. The research concentrates on forecasting if students will pass beyond the minimal 60% requirement because this threshold defines the campus recruitment process. The study uses a dataset obtained from second-year undergraduate students, including various academic and demographic features. Applying preprocessing and feature selection methods produces key academic achievement factors from the collected dataset. The developed machine learning model utilizes ABC and ACO optimization techniques to optimize the accuracy outcome. To further enhance user involvement and prediction capabilities, the work also entails establishing a web-based interface for the predictive model. The model's efficiency is validated by comparing it to the state-of-the-art models. The research highlights the significance of ML in analyzing educational data, demonstrating how it can enhance campus selection processes and increase accuracy in forecasting student achievement. Academic performance prediction is the primary focus of this study, although the methods and approaches used here have extensive potential for optimization of other selection criteria or assessment of student performance in various educational settings.

## 1.11 THESIS ORGANIZATION

**Chapter 1: Introduction**

This chapter summarizes the research on applying ML approaches to forecast students' academic performance. It explains the importance of predicting whether students will get the 60% threshold needed to be campus-selected. This chapter provides context for the study by discussing the research question, the importance of reliable performance prediction, and the methods used, such as feature selection, model development, and optimization strategies.

**Chapter 2: Literature Review**

This chapter conducts an all-inclusive examination of present-day research which focuses on predicting academic success and campus selection criteria. This research fills in current methodological gaps and includes past studies regarding machine learning models in education. The chapter reviews different prediction methods and techniques which focus on current scientific progress while identifying research gaps.

**Chapter 3: Research Methodology**

An extensive description of the research approach is detailed in this chapter regarding prediction model creation and evaluation. All phases of data acquisition, cleaning, feature optimization, and ML model development are included. The chapter explains web-based interface development for the model and covers optimization techniques.

**Chapter 4: Results Analysis**

The research examines the testing results and deployment of the predictive model in this chapter, while also analyzing both outcomes. The model's effectiveness is established through a complete performance assessment and comparison with existing models to prove its validity.

**Chapter 5: Discussion and Conclusion**

This chapter summarizes the key results obtained from the research, reflects on their implications, and discusses their effect on campus selection processes. The research recommends future exploration through which the prediction model can be enhanced, while also exploring other applications of ML in educational settings.

## Chapter 2

# LITERATURE REVIEW

This chapter reviews existing studies related to the prediction of student academic performance (SAP) using machine learning (ML) and optimization techniques. It explores the various factors influencing SAP, the algorithms commonly applied for prediction, and the recent advancements in developing web-based predictive systems. The review is organized into several sections, including an overview of SAP, Educational Data Mining (EDM), different ML and optimization approaches, as well as web-based interfaces, tools, and programming languages applied in prior research.

## 2.1 OVERVIEW OF STUDENT ACADEMIC PERFORMANCE (SAP)

Student Academic Performance (SAP) is a key indicator of educational success, influenced by multiple factors such as socio-economic background, learning habits, behavioral traits, and other related aspects [15]. Accurate prediction of academic outcomes enables early identification of at-risk students and supports timely interventions to improve learning outcomes [69]. In higher education, SAP prediction also plays an essential role in campus recruitment processes, where performance thresholds are often used to evaluate and select candidates.

## 2.2 EDUCATIONAL DATA MINING (EDM)

Educational Data Mining (EDM) is the process of analyzing educational data to determine important patterns and gain values that can enhance student learning outcomes [70]. It focuses on extracting meaningful information from large datasets in education to predict academic performance, detect learning issues, and recommend personalized interventions [24]. EDM empowers institutions to make data-driven decisions that improve student success [71], [72]. Its benefits extend to areas such as identifying key parameters for student performance evaluation, improving graduation rates, assessing teacher and student performance, generating reports, developing tools and interfaces, and effectively forecasting both management and student growth rates [70]. This section explores research studies across seven dimensions of EDM: decision support systems, adaptive learning, performance evaluation, text mining, outlier detection, scientific inquiry, and opinion mining.

**Decision Support Systems (DSS)** is developed to support administrators, teachers, and policymakers by analyzing large datasets and offering data-driven recommendations [73]. EDM, when integrated with DSS, can enhance decision-making in various areas such as student performance monitoring, resource allocation, curriculum development, and intervention strategies for at-risk students [74], [75].

[3] analysed data from 216 MBA students at CMS Business School, sourced from the Kaggle website. This study tested various ML algorithms to predict the appropriate undergraduate majors for students before their admission, considering current job market trends and their prior experiences. The researchers suggested the future implementation of intelligent recommender systems.

[45] utilized a first sample over four semesters from 2015 to 2017, involving 1,588 students, while the second sample included 1,282 students from the same period. This study employed random forest and logistic regression models to develop early warning systems for predicting student success and DSS. Various algorithms were tested; among them, LR and RF were selected due to their superior accuracy compared to the other methods.

[50] employed a dataset consisting of 270 entries gathered through online forms. The researchers developed a web application functioning as a DSS, aimed at helping students choose career paths that suit their personality traits, interests, and suitability for various courses. This application also provides recommendations for the most suitable colleges based on students' geographical locations and fee structures. The study applied various Decision Tree techniques, including ID3, CART, and CHAID, along with the Naive Bayes algorithm. Of these methods, CHAID achieved the highest accuracy of 40.09%, while ID3 recorded an accuracy of 66.37%, and C4.5 yielded 35.39%. Additionally, the Naive Bayes algorithm was assessed on 280 data points, resulting in an efficiency rate of 36.5%.

[51] analyzed a dataset of 12,477 students from Colombia and presented survey results to examine the academic decisions students face and the variables influencing these choices. The authors applied machine learning algorithms in a real-world case study to predict graduation rates. The researchers compared the performance of SVM and ANN, concluding that SVM demonstrated superior accuracy and was easier to implement than ANN.

[52] analyzed students enrolled in Mathematics and Knowledge Management courses during the 2016-2017 academic year. This study aimed to predict students' final scores before their final examinations by proposing an Early Recognition System that utilized real data captured from a blended learning course. The authors tested two algorithms, Neural Networks (NN) and Support Vector Machines (SVM), along with three hypotheses to determine the number of students at risk. The results indicated that NN performed better under certain conditions.

[76] conducted a study using data from 300 students across three colleges in Assam, India. The researchers utilized EDM tools and techniques within a DSS framework to enhance student performance and reduce dropout rates. The results demonstrated that RF outperformed the other classifiers. Additionally, the Apriori algorithm was employed within the DSS to discover association rules among the attributes, with the best rules being highlighted.

[77] analyzed data from 19 students, each assessed on 9 attributes related to scholarship eligibility. Key factors like family income, number of dependents, and academic as well as non-academic accomplishments were considered. Only students meeting these criteria were deemed eligible for scholarships. The dataset, collected via interviews, was processed using the C4.5 algorithm within the WEKA platform. Results showed that the Decision Support System (DSS) developed for evaluating scholarship recipients achieved a prediction accuracy of 94.74% for senior high school students' final grades.

[78] analyzed data from Australian students using three types of datasets, enrollment records, Moodle activity data, and a combination of both. The researchers used demographic and academic data captured during enrollment, along with participation

and assessment details from Moodle, to determine the factors impacting student performance. Various subgroup discovery techniques were employed, including SNS, DSSD, NMEEFSD (from the KEEL tool), BSD and SD-Map (both from Vikamine), and APRIORI-SD (from Orange). The results indicated that students from lower socio-economic backgrounds or those admitted under special entry criteria were at a higher risk of failing.

[79] analyzed data from 100 students enrolled in different distance learning platforms. The authors focused on applying knowledge discovery techniques to assess and forecast student dropout rates. They created a DSS that features a tool and a tutoring plan designed to mitigate dropout rates in e-learning courses. Logistic Regression models were utilized in this study, which helped institutions manage their revenue sources more effectively while improving student retention.

[80] analyzed data from 58,871 students over a span of seven years (2002-2008) in Peru. The authors discussed the application of a recommender system based on DSS. This system enables students to select the number and types of courses to enroll in, based on the experiences of previous cohorts with similar academic achievements. The C4.5 algorithm was applied to the entire dataset, resulting in the identification of effective patterns for the recommender system.

[81] focused on French and Lebanese students using university portals. The researchers proposed a recommender system. The primary goal was to identify students' needs, capabilities, preferences, and interests in order to recommend suitable majors and universities for each individual. While the authors outlined their approach, they indicated that implementation would take place in future work.

[82] presented different types of recommendation systems (RS) that varied particularly in terms of the recommendation algorithm used to predict the utility of a recommendation. Additionally, the authors discussed various challenges encountered while working in the RS domain and the role of DSS in enhancing these recommendations.

[83] utilized the UNSW-NB15 dataset, which comprised 2,540,044 records and 42

attributes. The authors proposed an effective intrusion detection system through a hybrid data optimization approach that included data sampling and feature selection, referred to as DO-IDS. Isolation Forest was employed to remove outliers, while a genetic algorithm optimized the sampling ratio. The Random Forest classifier served as the evaluation criterion for deriving the optimal training dataset. The implementation was carried out in Python within the PyCharm 2017 environment using Anaconda 3. The findings underscored the potential role of DSS in enhancing IDS performance.

[84] conducted three experiments using a total of 37 well-known benchmark problems to evaluate the performance of the proposed method. This paper introduced a hybrid algorithm that combines the genetic algorithm (GA), known for its strong global searching ability in DSS. A hybrid GAVNS algorithm, which integrates GA and VNS, was proposed. The results demonstrated that the proposed algorithm performed exceptionally well, discovering new solutions and outperforming other methods.

[85] evaluated data from Islamic Religious Education Learning in Turkey. This research described the planning and implementation of online-based exam evaluations in Learning. The research evaluated the three stages of the students' examination process, logging in, working on questions, and completing the test.

[86] examined how ML techniques could be utilized to automate the generation of reports and classify issue reports into bug and non-bug categories for DSS. The study applied various classification algorithms, including Naive Bayes, linear discriminant analysis, k-nearest neighbors, support vector machines, decision trees, and random forests (RF). The experiments revealed that RF outperformed the other algorithms and could be utilized in the decision support system process.

[87] reviewed 56 articles published between 2008 and 2018, focusing on two educational databases and one computer science database. This paper examined research studies based on recommender systems in education, as well as the methods used for evaluation and the results of those evaluations. The authors employed the Evidence for Policy and Practice Information (EPPI) approach to analyze the contexts

in which recommender systems are utilized.

[88] aimed to study 68 articles to understand the role that recommender systems play in assisting learners in developing self-regulation and decision-making skills. The authors employed the PRISMA protocol for the article selection process. Self-regulated learning (SRL) is widely utilized in online learning platforms, and the studies indicated that users generally responded positively to the recommendations for regulated learning.

[89] focused on young children as the target audience for recommender systems in education. The paper explores the complexities of RS in this field, aiming to articulate, label, and organize the specific layers of complexity observed. The authors reviewed various papers on RS and discussed the complex problems that may arise when using RS and decision support systems in education.

**Table 2.1. Summary of the reviewed literature on Decision Support Systems**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [3] | 2021 | Analyzed data from 216 MBA students at CMS Business School to predict undergraduate majors based on job trends and prior experiences. | Gradient Boosting Classifier (GBC) and RF. | RF and GBC were suitable for integration into intelligent recommender systems and RF achieved highest accuracy of 75%. | Lacked detailed implementation of an intelligent recommender system. |
| [45] | 2019 | Collected data from two samples to develop early warning systems for predicting success and DSS. | Logistic Regression (LR), KNN, CART, NB, SVM, RF. | LR and RF demonstrated superior accuracy compared to other tested algorithms and LR achieved highest accuracy of 80%. | Did not analyze longitudinal data to track long-term student outcomes. |

| | | | | | |
|---|---|---|---|---|---|
| [50] | 2020 | Developed a web-based DSS to help 270 students choose career paths and colleges based on personality traits, interests, and geography. | Decision Tree algorithms (ID3, CART, CHAID, C4.5) and Naïve Bayes. | ID3 achieved the highest accuracy (66.37%), while CHAID and C4.5 recorded lower accuracy levels. | Low prediction accuracy with some algorithms and limited dataset size. |
| [51] | 2018 | Analyzed data from 12,477 students in Colombia to predict graduation rates and support academic decision-making. | ANN and SVM. | SVM demonstrated superior accuracy and was easier to implement than ANN and SVM achieved highest accuracy of 82%. | Lack of detailed analysis on factors influencing algorithm performance. |
| [52] | 2018 | Investigated students' final scores in Mathematics and Knowledge Management courses to develop an Early Recognition System for identifying at-risk students. | Neural Networks (NN), SVM. | NN outperformed SVM under certain conditions for predicting student final scores and SVM achieved highest accuracy of 75%. | Limited to a single academic year and two courses. |
| [76] | 2018 | Conducted a study on 300 students across three colleges in Assam, India, to enhance student performance and reduce dropouts. | J48, PART, Random Forest (RF), Bayes Network, Apriori algorithm (WEKA). | RF outperformed other classifiers; Apriori discovered association rules among attributes and RF achieved highest accuracy of 84.3%. | Focused only on a small dataset from a limited geographical region. |

| | | | | | |
|---|---|---|---|---|---|
| [77] | 2018 | Assessed 19 students on 9 attributes related to scholarship eligibility, including family income, dependents, and accomplishments. | C4.5 algorithm (WEKA). | DSS achieved 94.74% prediction accuracy for senior high school students' final grades. | Dataset was small and limited to specific eligibility attributes. |
| [78] | 2020 | Analyzed Australian students' performance using enrollment records, Moodle activity data, and combined datasets to determine factors affecting performance. | Subgroup discovery techniques (SNS, DSSD, NMEEFSD, BSD, SD-Map, APRIORI-SD). | Students from lower socio-economic backgrounds or special entry criteria were at higher risk of failing and SNS achieved highest accuracy of 93.3%. | Did not explore intervention strategies for at-risk students identified. |
| [79] | 2017 | Analyzed data from 100 students enrolled in different distance learning platforms. The study focused on applying knowledge discovery techniques to assess and forecast student dropout rates. | Logistic Regression. | Developed a DSS to forecast student dropout rates and improve retention and achieved accuracy of 97.13%. | Lack of specific application to different disciplines. |
| [80] | 2009 | Analyzed data from 58,871 students over a span of seven years (2002-2008) in Peru. The research applied a RS to assist students in making informed academic decisions. | C4.5 Algorithm. | Identified effective patterns for course selection based on past cohorts achieved accuracy of 81.38%. | Limited exploration of factors beyond previous academic performance. |

| | | | | | |
|---|---|---|---|---|---|
| [81] | 2018 | Focused on French and Lebanese students using university portals. The study proposed an ontology-based recommender system enhanced with machine learning techniques. | K-mode, Self-organizing maps, Hierarchical clustering. | Designed a system to recommend suitable majors and universities based on students' needs and preferences. | Implementation of the proposed system is not yet conducted. |
| [82] | 2015 | Presented various types of recommendation systems (RS) in different domains, focusing on the recommendation algorithm in DSS. | Various recommendation algorithms. | Discussed the role of DSS in enhancing recommendations, especially in educational domains. | The study lacks specific application to educational settings. |
| [83] | 2019 | Utilized the UNSW-NB15 dataset (2,540,044 records) for intrusion detection system (IDS) optimization using hybrid data sampling and feature selection. | Isolation Forest, Genetic Algorithm, Random Forest. | DSS improved IDS performance through optimized data sampling and feature selection by 93.4% accuracy. | Study is limited to IDS and not directly applicable to student academic performance. |
| [84] | 2018 | Conducted three experiments on 37 benchmark problems to evaluate the performance of a hybrid genetic algorithm (GA) and variable neighborhood search (VNS) for integrated planning and scheduling. | Hybrid GAVNS Algorithm. | The hybrid algorithm outperformed (79.2%) other methods, discovering new solutions for scheduling problems. | Focused on planning and scheduling problems, not educational contexts. |

| | | | | | |
|---|---|---|---|---|---|
| [85] | 2022 | Evaluated data from Islamic Religious Education Learning in Turkey. A qualitative case study assessed online-based exam evaluations. | Observation, Interviews, Documentation. | Evaluated three stages of the online exam process for students in Islamic education. | Limited to a specific case study in religious education. |
| [86] | 2017 | Examined the use of ML techniques to automate report generation and classify issues into bug and non-bug categories in decision support systems. | Naive Bayes, Linear Discriminant Analysis, kNN, SVM, Decision Trees, Random Forest. | RF outperformed with 86.2% accuracy as compared to other algorithms for classifying bug reports and automating decision support. | Limited to bug classification, not directly applicable to academic decision-making. |
| [87] | 2020 | Reviewed 56 articles on recommender systems in education, focusing on evaluation methods and results. | Evidence for Policy and Practice Information (EPPI). | Identified key trends and methods in recommender systems for education. | Lack of focus on newer technologies and applications. |
| [88] | 2021 | Studied 68 articles to understand the role of recommender systems in self-regulation and decision-making for learners. | PRISMA Protocol. | Found that learners responded positively to recommendations promoting self-regulation in online learning. | Limited application to specific types of learning platforms. |
| [89] | 2019 | Focused on young children and explored complexities in using recommender systems in education. | Review of Various RS Techniques. | Identified complexities and challenges in applying RS to young learners. | Lack of implementation of proposed solutions in practical scenarios. |

**Adaptive Learning** involves tailoring educational experiences to meet the personal needs of students. By determining data gathered from various learning activities, adaptive learning systems can identify a learner's strengths, weaknesses, preferences, and learning styles [90], [91]. This personalized approach enables educators to provide targeted interventions, resources, and assessments, enhancing student engagement and improving learning outcomes [41].

[92] studied 77 Computer Science students in Saudi Arabia. The study examined the effect of adaptive and traditional gamification on the effectiveness of e-learning. An Adaptive Gamification Learning System app was also proposed by the researchers. The findings indicated that students using adaptive gamification outperformed those using traditional gamification and demonstrated significantly better performance.

**Table 2.2. Summary of the reviewed literature on Adaptive Learning**

| **Ref.** | **Year** | **Study Description** | **Method/ Algorithm Used** | **Key Findings** | **Gap** |
|---|---|---|---|---|---|
| [92] | 2020 | Studied 77 Computer Science students in Saudi Arabia to examine the effect of adaptive and traditional gamification on the effectiveness of e-learning. | Adaptive Gamification Learning System (DT-Gini index and NB). | Students using adaptive gamification outperformed those using traditional gamification, demonstrating higher engagement and improved learning performance (DT with 92% accuracy). . | Limited to a small sample size. |

**Performance Evaluation** in EDM involves assessing the effectiveness of predictive models and algorithms used to analyze student data [93]. The evaluations help optimize models for better decision-making in educational environments, ensuring the reliability and applicability of data-driven insights [94].

[94] studied the examination marks of 72 undergraduate students across four registered courses. The research focused on evaluating the uses of data mining techniques in higher education, particularly consolidating clustering algorithms in the context of Educational Data Mining. The use of K-means clustering helped institutions monitor

students' performance semester by semester, with results being used to identify measures for improving academic outcomes in future semesters.

[95] studied 213 students from various courses at a public university in Jeddah. The paper investigated machine learning techniques to forecast course grades for these students. The primary algorithm, KNN with Z-score, was compared with other models, including BaselineOnly, SlopeOne, Normal Predictor, Non-negative Matrix Factorization (NMF), CoClustering, Singular Value Decomposition (SVD), and KNNBasic. The results proved that the KNN with Z-score model outperformed the others in decision-making tasks.

[96] reviewed papers from 2009 to 2019, focusing on the use of alternative ML methods. The study also explored the combination of SVM with Partial Least Squares Discriminant Analysis (PLS-DA). The authors concluded that, while SVM is effective, conventional methods like PLS-DA remain highly relevant and impactful in this field.

**Table 2.3. Summary of the reviewed literature on Performance Evaluation**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [94] | 2018 | Studied examination marks of 72 undergraduate students across four registered courses to evaluate the use of data mining techniques in monitoring academic performance. | K-Means Clustering. | K-Means Clustering enabled institutions to monitor students' performance semester by semester and identify measures to improve future academic outcomes with 80.3% accuracy. | Limited to clustering techniques; lacks predictive modeling. |
| [95] | 2015 | Studied 213 students from various courses at a public university in Jeddah to forecast course grades using machine learning techniques. | KNN with Z-score, BaselineOnly, SlopeOne, Normal Predictor, NMF, CoClustering, SVD, and KNNBasic. | KNN with Z-score outperformed (84%) other models in forecasting course grades and decision-making tasks. | Limited comparison of feature selection techniques. |

| | | | | | |
|---|---|---|---|---|---|
| [96] | 2019 | Reviewed research papers (2009-2019) on the application of ML methods for food analysis evaluation. | SVM, CART, Random Forest, PLS-DA. | SVM is effective for food analysis, but conventional methods like PLS-DA remain highly relevant and impactful. | Limited focus on hybrid ML approaches. |

**Text Mining** techniques are developed to assist educators, administrators, and researchers by extracting valuable insights from large volumes of unstructured data [97], [98]. When integrated with educational environments, text mining can improve decision-making processes in areas such as student feedback analysis, curriculum enhancement [99], identifying learning patterns, and predicting academic outcomes for better intervention strategies [100].

[97] analyzed data from students, teachers, employers, and graduates at a university in Spain. The research aimed to evaluate methods based on text network and topic modeling to extract important insights from open ended questions. The surveys sought to assess the level of satisfaction with academic processes at the university. The study focused on the use of clustering algorithms and employed Latent Dirichlet Allocation for topic modeling.

[98] examined 66 thousand student reviews for MOOCs to develop an effective sentiment classification framework using text mining, deep learning techniuqes, ensemble learning, and ML techniques. The research employed five ML algorithms, SVM, Naive Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Random Forest (RF). Additionally, five ensemble learning approaches were utilized, including Bagging, AdaBoost, Stacked Generalization, Voting, and Random Subspace. The study also applied five deep learning architectures, and the findings indicated that ensemble methods outperformed traditional supervised learning algorithms in the context of EDM.

[101] analyzed data from a computer science course at North Texas University, which included 29 students' submissions for 10 assignments and two exams. A predictive model was developed to estimate student scores by assessing the similarity in student

responses and the provided model results. The findings suggested that computational methods could be used to enhance the consistency of human grading, rather than replacing it. Techniques such as Linear Regression, K-Means clustering, and the Bag of Words were applied. The approach proved effective, particularly when a well-defined vocabulary for the model answers was available.

[102] analyzed 430 web documents from 172 websites, introducing a text mining approach to extract expert information from the Internet. The framework, along with several modules and algorithms, was developed to support this process. An expert recommendation model and a weighted directed graph were key components, with Python and MySQL used for implementation. The results demonstrated that the proposed text mining algorithm accurately extracted expert data, making it valuable for student assessment processes.

[103] collected data from Ilan University spanning the 2014 to 2016 academic years. This study employed sentiment analysis to evaluate students' written feedback. The analysis was conducted using the Chinese-text sentiment analysis toolkit. The results highlighted that the LSTM classifier was the most effective, achieving a 97% accuracy rate for identifying positive sentiments and an 87% accuracy rate for negative sentiments.

[104] collected data comprised opinions from over 2,600 teachers who participated in e-learning courses. A substantial amount of feedback from learners attending lifelong e-learning courses was gathered through questionnaires. Various algorithms, including SVM, RF, NB, and J48, were utilized for testing and training the data. The RF algorithm achieved the best performance at 63.15%.

[105] analyzed a total of 343 papers published between January 2006 and July 2018 from various databases. The study aimed to identify the most commonly utilized text mining techniques, the frequently employed educational resources, and the primary applications or objectives within educational settings. A systematic review was conducted to compile all published works on the use of text mining in education.

[106] analyzed 300 research articles in mobile learning published between 1990 and

2016. The study aimed to employ text mining techniques to identify topics within scientific texts and establish hierarchical and evolutionary connections among these topics, while also utilizing visualization tools. The K-means algorithm was applied with different k-values to create six clusters using the RapidMiner tool. Results indicated that the similarity operator was unable to identify clear similarities among some topics due to their interrelatedness and similar meanings in educational data mining.

**Table 2.4. Summary of the reviewed literature on Text Mining**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [97] | 2020 | Analyzed data from students, teachers, employers, and graduates at a university in Spain to evaluate satisfaction with academic processes. | Topic modeling (Latent Dirichlet Allocation), text network modeling, clustering algorithms. | Outlined techniques employed by educators to enhance student retention. | Limited to surveys; may not generalize to other types of academic data. |
| [98] | 2021 | Examined 66,000 student reviews for MOOCs to develop a sentiment classification framework. | Text mining, five ML algorithms (SVM, NB, KNN, LR, RF), five ensemble learning methods, and five deep learning architectures (RNN, LSTM, CNN, GRU). | KNN with Z-score outperformed (91.1%) other models in forecasting course grades and decision-making tasks. | Limited comparison of feature selection techniques. |
| [101] | 2019 | Analyzed data from a computer science course to predict student scores. | Linear Regression, K-Means clustering, Bag of Words. | Suggested computational methods could enhance grading consistency without replacing human grading with 73.5% of accuracy of LR. | Relied on a well-defined vocabulary for effectiveness. |

| | | | | | |
|---|---|---|---|---|---|
| [102] | 2019 | Analyzed 430 web documents from 172 websites to extract expert information. | Text mining, expert recommendation model, weighted directed graph, Python, MySQL. | Accurately extracted expert data, valuable for student assessment processes. | Limited scalability and context-specific applications. |
| [103] | 2018 | Collected student feedback data from Ilan University for sentiment analysis. | SnowNLP toolkit, NB, RNN, LSTM. | LSTM achieved the highest accuracy (97% positive, 87% negative sentiments). | Limited to Chinese text; applicability to other languages untested. |
| [104] | 2018 | Collected feedback from over 2,600 teachers and learners in e-learning courses. | SVM, RF, NB, J48. | RF achieved the best performance (63.15%). | Limited accuracy achieved compared to advanced models. |
| [105] | 2019 | Reviewed 343 papers (2006–2018) to identify text mining techniques, educational resources, and applications. | Systematic review of text mining techniques in education. | Identified commonly used techniques and applications in education. | Limited exploration of recent advancements in text mining. |
| [106] | 2018 | Analyzed 300 articles on mobile learning (1990–2016) to identify topics and hierarchical connections. | Text mining, K-Means clustering, visualization tools (RapidMiner). | Created six clusters, but some topics lacked clear similarities due to interrelatedness. | Difficulty in differentiating similar topics in educational data mining. |

**Outlier Detection** in EDM is the process of identifying data points that significantly deviate from the majority of the dataset [55]. These outliers can affect model performance and may indicate errors, rare events, or novel patterns [109]. Common detection methods include statistical techniques, distance- or density-based approaches, and machine learning algorithms like Isolation Forest or One-Class SVM.

Detecting outliers is crucial for improving model accuracy and uncovering valuable insights [113].

[10] collected data from 273 registered students in the UK. The study examined what could be considered outliers in time-on-task, which varied based on each student's teaching-learning pattern, the stages of the learning process, and the nature of the tasks involved. The researchers utilized Kovanovic's approach, and all data analyses were conducted using R v3.5.3. The results indicated that using an arbitrary cut-off value could potentially lead to an underestimation of outliers.

[55] proposed a method for improving data quality by detecting incorrect data based on an enhanced Local Outlier Factor (LOF). This method effectively identified both missing and abnormal segments, two common types of incorrect data providing valuable insights for cleaning big data in machinery condition monitoring.

[107] utilized Prestige data from R for multivariate analysis. The study examined tools for detecting univariate outliers, focusing on Mahalanobis distance and multivariate statistical distances to identify outliers. The findings indicated that Mahalanobis distance was more suitable for handling correlated multivariate data compared to Euclidean distance. However, further examination of Mahalanobis distance using diverse datasets and updated methodologies was suggested.

[108] introduced a stream mining algorithm to cluster data streams and track their evolution using DBSCAN. The method operated on three principles: Build Clusters, Merge, and Prune. The proposed approach effectively detected outliers, analyzed cluster evolution, and examined the relationship between outlier detection and physical events.

[109] utilized data from Samsung in Saudi Arabia to propose a new Dixon's test for identifying outliers in datasets, reforming it with fuzzy logic. This integration of statistical methods with fuzzy logic improved the accuracy of Dixon's test compared to previous methods.

[110] collected 103 samples using the simple random sampling method to analyze non-

academic factors that influenced student graduation. The study aimed to create a predictive model with high accuracy by employing classification methods such as KNN, SVM, and DT. The DT model demonstrated the highest accuracy with a training time of 1.7 seconds.

[111] proposed a weighted-outlier-mining method named WATCH to determine outliers, utilizing eight real-world datasets. The GAClust toolkit was employed to generate synthetic datasets, and precision measures were used to quantify the accurcay of the outlier detection algorithms. When compared to existing algorithms such as Incremental Tree-Based (ITB), Adaptive Voting Framework (AVF), and Genetic Algorithm (GA), the results indicated that the feature grouping implemented in WATCH outperformed the others.

[112] involved 520 participants from Canada and proposed a data-quality evaluation process that focused to identify errors, establishing an iterative framework. The study utilized the ML techniques in R, demonstrating that these two algorithms were effective in detecting outliers.

[113] utilized a UCI Repository dataset and described a threshold-based approach for extracting significant features, asserting that only attribute values and their corresponding information gain are necessary. This method, implemented using MATLAB R2017a, proved effective in managing outliers while identifying students at risk.

[114] utilized 18 real datasets to create an unsupervised Outlying Feature Value Set (OUVAS), aimed at developing a high-quality set for detecting outliers in noisy categorical data through hierarchical clustering. The findings indicated that the Random walk-based Hierarchical Value Couplings (RHAC) outlier detection method, along with an Extended RHAC-based feature selection approach, improved the performance of traditional outlier detectors.

**Table 2.5. Summary of the reviewed literature on Outlier Detection**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| [10] | 2020 | Collected data from 273 registered students in the UK to examine outliers in time-on-task based on learning patterns, task stages, and task nature. | Kovanovic's approach; R v3.5.3. | Using an arbitrary cut-off value could lead to underestimation of outliers. | Limited to a single analytical approach for outlier detection. |
| [55] | 2019 | Proposed a method to improve data quality by detecting incorrect data using an enhanced LOF approach. | Enhanced Local Outlier Factor (LOF). | Effectively identified missing and abnormal data, improving big data cleaning in machinery condition monitoring. | Focused primarily on machinery condition monitoring. |
| [107] | 2019 | Utilized Prestige data for multivariate analysis to detect univariate outliers using statistical distances. | Mahalanobis distance, multivariate statistical distances. | Mahalanobis distance (83.2%) outperformed Euclidean distance for correlated multivariate data. | Limited examination across diverse datasets. |
| [108] | 2021 | Introduced a stream mining algorithm to cluster data streams and detect outliers while tracking cluster evolution. | DBSCAN. | The approach effectively detected outliers, analyzed cluster evolution, and examined relationships with physical events with 79.4% accuracy. | Limited to stream data and specific event tracking. |
| [109] | 2020 | Utilized Samsung data to improve Dixon's test for outlier detection by integrating fuzzy logic. | Dixon's test with fuzzy logic. | Improved accuracy of Dixon’s test compared to previous methods. | Limited focus on fuzzy logic integration. |

| | | | | | |
|---|---|---|---|---|---|
| [110] | 2018 | Collected 103 samples to analyze non-academic factors influencing student graduation using classification models. | KNN, SVM, Decision Tree (DT). | DT demonstrated the highest accuracy of 66% with a training time of 1.7 seconds. | Limited dataset and non-academic factors only. |
| [111] | 2018 | Proposed a weighted-outlier-mining method, to determine outliers. | WATCH, GAClust toolkit, Incremental Tree-Based (ITB), AVF, GA. | Feature grouping in WATCH outperformed existing algorithms. | Limited to high-dimensional categorical data. |
| [112] | 2019 | Involved 520 participants to propose a data-quality-evaluation process. | Minimum Covariance Determinant (MCD), RPCA; R. | MCD and RPCA effectively detected multivariate outliers. | Limited to specific multivariate methods and frameworks. |
| [113] | 2019 | Used a UCI Repository dataset to describe a threshold-based approach for extracting significant features and managing outliers. | Threshold-based feature extraction, MATLAB R2017a. | Effective in identifying at-risk students while managing outliers. | Limited to threshold-based techniques. |
| [114] | 2022 | Developed OUVAS for detecting outliers in noisy categorical data through hierarchical clustering. | RHAC, Extended RHAC-based feature selection. | Improved performance of outlier detection (69.3%) and outperformed contemporary feature selection methods. | Focused on noisy categorical data only. |

**Scientific Inquiry** involves students identifying research problems and finding solutions. The increasing impact of ML on real-world scientific investigations demands a shift to teaching from traditional method to scientific inquiry ML methods [115].

[115] explored three key questions regarding the modeling of inquiry-based learning to integrate scientific inquiry into daily teaching practices, the benefits of using stories to support the inquiry process, and practical examples of employing stories to teach scientific inquiry skills. It presented a framework aligned with England's curriculum, demonstrating that the system effectively guided early years and primary teachers in embedding scientific inquiry skills and knowledge.

**Table 2.6. Summary of the reviewed literature on Scientific Inquiry**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [115] | 2020 | Explored three questions: integrating scientific inquiry into teaching practices, benefits of using stories to support inquiry, and examples of employing stories to teach inquiry skills. | Framework for integrating inquiry-based learning with stories. | Demonstrated that the framework effectively guided teachers in embedding scientific inquiry skills, enhancing students' understanding of the world. | Focused primarily on early years and primary education in the context of England's curriculum. |

**Opinion Mining**, also known as sentiment analysis involves identifying the sentiment of a statement, which can then be analyzed at various levels, such as paragraphs, documents, sentences, and texts [36]. It uses various computational techniques to extract, process, and analyze text data [116], [118].

[36] collected student feedback data from module evaluations in Oman, comprising a total of 6,433 responses. The study conducted opinion mining on this feedback using the NB and KNN supervised ML algorithms implemented through RapidMiner. The results indicated that NB was the most accurate, while KNN excelled in precision.

[116] gathered feedback from 120 third-year students in Malaysia to develop a system called OMFeedback that could automatically analyze student feedback. The system employed a lexicon-based approach for sentiment analysis, capable of interpreting capitalization and emoji characters in the feedback.

[117] collected data from a reviewed paper to present a strategy for using opinion mining to compute the sentiment polarity of student comments, illustrated through a toy example. The analysis was based on open-ended survey questions, and the authors concluded that further exploration of opinion mining is needed for larger datasets.

[118] developed an Online Teacher's Evaluation Tool that can be utilized by universities to evaluate teaching performance. The primary goal was to assign specific weight value to narrative responses from students as part of the performance rating scale. As a result, the tool can generate reports in PDF format after evaluation.

[119] collected data from the Department of Languages at Mandalay, noting that previous studies did not display the level of opinion results as positive or negative. To address this, the researchers proposed an automatic analysis of student text feedback using a lexicon-based approach to predict teaching performance levels.

**Table 2.7. Summary of the reviewed literature on Opinion Mining**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [36] | 2016 | Collected student feedback data from module evaluations in Oman (6,433 responses). | KNN and NB using RapidMiner. | NB was the most accurate algorithm (87.3%), while KNN showed better precision. | Limited to two algorithms (NB and KNN), without exploring other advanced ML techniques. |
| [116] | 2019 | Developed OMFeedback, a system for automatic sentiment analysis of feedback from 120 third-year students in Malaysia. | Lexicon-based approach for sentiment analysis. | Successfully analyzed sentiment, including capitalized words and emojis, with meaningful results with 60.3% accuracy. | Focused only on lexicon-based analysis; lacked integration with supervised ML or deep learning. |

| | | | | | |
|---|---|---|---|---|---|
| [117] | 2021 | Proposed a strategy to compute sentiment polarity of student comments using a toy example. | Sentiment analysis of open-ended survey responses. | Concluded that larger datasets require further exploration for opinion mining to be effective. | Demonstrated analysis only on a small dataset with limited generalizability. |
| [118] | 2016 | Developed an Online Teacher's Evaluation Tool for analyzing teaching performance. | Weighting narrative responses in performance analysis. | Successfully combined qualitative and quantitative ratings, generating PDF reports. | Focused only on evaluation tools and lacked predictive modeling. |
| [119] | 2017 | Collected feedback from the Department of Languages at Mandalay to predict teaching performance levels. | Lexicon-based sentiment analysis. | Proposed a method for predicting teaching performance levels with 85% accuracy. | Lack of comparative analysis with other sentiment analysis approaches. |

## 2.3 ARTIFICIAL INTELLIGENCE IN EDUCATION (AIEd)

Artificial Intelligence (AI) plays a transformative contribution in education by enabling intelligent tutoring systems, adaptive learning platforms, and automated grading [120]. AI-driven systems analyze student behaviors and learning patterns to provide personalized recommendations, predict academic success, and enhance overall learning efficiency. AI tools are essential for automating processes and supporting personalized learning at scale.

[121] reviewed 69 articles (2018-2023) to evaluate challenges in adopting AI in education. The study recommended ChatGPT as a intelligent teaching tool, emphasizing the need for customization for educators. Key discussions included creative prompts, dataset diversity, bias elimination, and data confidentiality. Strategic solutions were proposed to enhance AI adoption in education.

[122] analyzed 2,223 research articles and reviewed 125 papers to explore AI

applications in education, such as adaptive learning, personalized tutoring, and intelligent assessment. The study examined key research topics, methodologies, and guiding theories while identifying challenges and underexplored areas. It provided insights to guide future research in the dynamic field of AIEd.

[123] conducted a case study in Taiwan to examine faculty perspectives and roles in higher education internationalization through in-depth interviews and participant observation at four universities. The findings revealed that while faculty supported internationalization, their strategies and responses varied. Internationalization influenced their work through EMI, performance-based evaluations, and a focus on international publications. Additionally, the study highlighted challenges faced by non-English-speaking universities competing globally.

[124] explored the use of AI for educational purposes to identify meaningful patterns. The study highlights how data, such as e-learning logs, demographic and academic student data, and registration information, can be leveraged by ML algorithms. The authors presented a case study predicting student marks and performance in written assignments as a training set for a regression model. Additionally, a prototype software tool was developed to assist tutors.

[125] deployed an AI prediction model with learning-analytics (LA) to enhance student learning outcomes in collaborative learning settings. The study addressed the gap where most AI models focus on algorithmic accuracy rather than offering continuous, timely feedback to students. A quasi-experimental study was conducted in an online engineering course, and the results demonstrated that this approach led to greater student engagement and increased satisfaction.

[126] conducted a review of 45 studies on AIEd. The analysis focused on aspects such as publication trends, leading journals, and the technologies used in AIEd research. The review found growing interest in AIEd but noted the limited use of advanced techniques like deep learning, with traditional AI methods like natural language processing being more common. The study recommended exploring AI's potential in physical classrooms, adopting advanced deep learning algorithms, and integrating AI

more closely with educational theories.

**Table 2.8. Summary of the reviewed literature on AIEd**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [121] | 2024 | Reviewed 69 articles (2018-2023) to evaluate challenges in adopting AI in education. | Categorization into user, operational, environmental, technological, and ethical dimensions. | Discussed ChatGPT creative prompts, dataset diversity, bias elimination, and data confidentiality. | Lack of standardization in AI adoption methods. |
| [122] | 2024 | Analyzed 2,223 articles and reviewed 125 papers to explore AI applications in education. | AI applications in adaptive learning, personalized tutoring, and intelligent assessment. | Identified research topics, methodologies, guiding theories, and underexplored areas. | Limited integration of AI in educational practices. |
| [123] | 2024 | Case study in Taiwan examining faculty perspectives on internationalization in higher education. | In-depth interviews and participant observation at four universities. | Faculty supported internationalization, but strategies and responses varied. Identified challenges faced by non-English-speaking universities. | Lack of consistency in faculty strategies. |
| [124] | 2012 | Explored AI in education for identifying meaningful patterns. | ML algorithms for predictive modeling. | Developed a case-study for predicting student marks. | Limited use of data in practical teaching. |
| [125] | 2023 | Deployed an AI prediction model with learning-analytics to enhance student learning outcomes. | LA and AL model. | Results showed higher student engagement, improved collaborative performance, and increased satisfaction with 88.6% accuracy. | Most AI models focus on algorithmic accuracy rather than feedback. |
| [126] | 2020 | Conducted a review of 45 state-of-the-art literature on AIEd. | AI techniques in educational contexts. | Growing interest in AIEd but limited use of DL techniques. | Lack of integration of AI with educational theories. |

## 2.4 DEEP LEARNING IN EDUCATION (DLEd)

Deep Learning (DL), a subset of AI, uses neural networks to model complex patterns in student data. DL models are effective in handling unstructured data, such as text and images, and are applied to analyse student essays, detect learning disabilities, and predict performance [127],[128].

[24] developed an embedded DL model to forecast student academic performance. The study focused on addressing the challenges of nonlinear relationships among factors influencing academic performance. A real-world student dataset was utilized for training and testing. Results showed that the hybrid model outperformed in accuracy and robustness, making it effective for early intervention and academic pre-warning systems.

[47] addressed the challenge of class imbalance in supervised learning, especially with sequential data. The study proposed a deep learning-based method using Sequential Conditional Generative Adversarial Networks (SC-GAN). SC-GAN captures past student behavior and generates synthetic data to mitigate class imbalance. The method showed superior performance compared to traditional up-sampling techniques, with improvements of 7.07% and 6.53% in AUC, respectively.

[98] investigated the application of sentiment analysis in Massive Open Online Courses using ML, ensemble learning, and DL approaches. The findings revealed that DL architectures outperformed both supervised and ensemble methods. Among the tested configurations, the GloVe word-embedding scheme achieved the highest classification accuracy of 95.80%.

[127] investigated the use of big data and AI to predict dropout rates among 3,552 university students in Taiwan. Key predictors included class ranking dynamics, student loan applications, school absences, and the number of alerted subjects. The study compared a statistical learning method, which achieved 68% accuracy, with a deep learning method, which reached 77%. Both methods achieved over 80% accuracy when specificity was prioritized. The findings demonstrated AI's potential to assist universities in identifying at-risk students early, enabling targeted interventions to

reduce dropouts, support adaptive learning, and improve student outcomes.

[128] explored predicting the binding modes of flexible polypeptides to proteins, a complex task in molecular docking. The study aimed to enhance the accuracy of Glide's docking tool for flexible polypeptides, focusing on improving sampling techniques. The findings demonstrated a success rate increase in docking accuracy from 21% to 58% when redocking to native protein structures, making the method both efficient and accurate. This improvement highlights the potential application of advanced docking techniques in predicting molecular interactions effectively, which can also support educational research in biochemistry and related fields.

[129] conducted a bibliometric knowledge synthesis study to address the small data problem in machine learning. Despite living in a 'Big Data' world, many real-world scenarios still involve small data samples, posing challenges for machine learning.

[130] focused on predicting anxiety disorder and depressive disorder using ML methods. The study involved 4,184 undergraduate students who underwent general health screening and psychiatric assessments. A machine learning pipeline, combining distinct algorithms including deep learning, was used, with 59 biomedical and demographic features along with engineered features for model training. The model achieved an AUC of 0.73 for GAD and 0.67 for MDD. SHAP values were used to identify key predictive features: satisfaction with living conditions and public health insurance for MDD, and vaccination status and marijuana use for GAD.

**Table 2.9. Summary of the reviewed literature on DLEd**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [24] | 2015 | Developed a hybrid deep learning model to predict student academic performance by combining CNNs and LSTMs. | LSTM and CNN. | Embedded model (77.2%) outperformed traditional methods in accuracy and robustness for early intervention and academic pre-warning systems. | Need for further optimization in model interpretability. |

| | | | | | |
|---|---|---|---|---|---|
| [47] | 2020 | Addressed the challenge of class-imbalance in supervised learning. | SC-GAN model. | SC-GAN improved AUC by 7.07% and 6.53% over Random Over-sampling and SMOTE. | Limitation in handling very complex data patterns. |
| [98] | 2021 | Investigated sentiment analysis in MOOCs using ML, ensemble learning, and DL. | GloVe word-embedding scheme and LSTM. | Deep learning outperformed other methods with 95.80% accuracy in sentiment analysis of MOOC reviews. | Exploration of different DL architectures in specific course contexts. |
| [127] | 2020 | Used AI and big data to predict dropout rates among university students in Taiwan. | Statistical Learning, Deep Learning | Deep learning achieved 77% accuracy, and over 80% accuracy when prioritizing specificity, identifying at-risk students early. | More accurate predictions with additional data features. |
| [128] | 2019 | Focused on predicting the binding modes of flexible polypeptides to proteins, improving docking accuracy. | Glide Docking Tool, Sampling Enhancement Techniques | Success rate in docking increased from 21% to 58%, improving efficiency and accuracy. | Potential for further improvements in flexibility handling. |
| [129] | 2022 | Addressed the small data problem in machine learning through bibliometric analysis. | Bibliometric Knowledge Synthesis | Identified growing research trends and key themes like data augmentation and statistical learning on small datasets. | Limited exploration of real-world applications for small data problems. |
| [130] | 2021 | Predicted anxiety disorder and depressive disorder using EHRs and machine learning. | Machine Learning and Deep Learning | AUC of 0.73 for GAD and 0.67 for MDD; key features include satisfaction with living conditions and vaccination status. | Limited data sources for predicting mental health conditions. |

## 2.5 MULTIFACETED FACTORS AFFECTING STUDENT ACADEMIC PERFORMANCE

SAP is affected by a variety of interconnected factors, each contributing uniquely to their educational outcomes. These factors span across multiple domains, including demographic, educational, behavioral, time-management, contextual, socioeconomic, and family-related aspects. Understanding these factors comprehensively is essential for developing an accurate predictive model for academic performance. The interplay of these factors shapes students' learning experiences and trajectories, ultimately affecting their academic success, well-being, and future prospects. Various researchers have explored these multifaceted factors, emphasizing their importance in shaping academic outcomes. The features under these multifaceted factors are depicted in Table 2.10, providing a comprehensive categorization of the features influencing student academic performance.

**Table 2.10. Feature Categorization Under Multifaceted Factors Influencing Student Academic Performance**

| Factors | Feature Description | Factors | Feature Description |
|---|---|---|---|
| Demographic Factors | Gender [13], [36], [128], [131], [133] | Family Factors | Single Parent Child[128], [131] |
| | Age[13], [128], [131] | | First Child[131] |
| | Marital Status[95] | | Family Support[28], [139] |
| | Number of Children[95] | | Family Conflict[128] |
| | | | Family Freedom[128] |
| | Religion[132] | | Family Moral Values[139] |
| | Category[77] | | Family Help, Trust[77] |
| | | | Family Type [128], [131] |
| Educational Factors | Co-Curricular Activities in College[77], [139], [140], [141] | | Number of Siblings [128], [131] |
| | | | Parents Martial Status [128], [131] |
| | Attendance[127], [139], [142], [167], [175] | Socioeconomic Factors | Part-Time Job [127], [139], [169] |
| | Father Academic Qualification[131], [139], [169] | | Total Income[77], [139], [140], [141] |
| | Mother Academic Qualification [131], [139], [169] | | Scholarship[77] |

| | | | |
|---|---|---|---|
| | Course Type[13], [77], [95], [141], [168] | | Mother Occupation[139] |
| | | | Father Occupation[139] |
| | Last year grade [77], [95], [127], [131], [139], [140], [141], [142], [143], [144], [145], [146], [147], [148], [149], [150], [151], [152] | Contextual Factors | Distance From College[95], [131], [141] |
| | Relationship with Classmates[36], [135] | | Effect of COVID-19 Pandemic[13], [170], [171], [172] |
| | Relationship with Teachers[36], [135], [162] | | Gender Inequality[13], [127], [128], [131], [167], [168], [169] |
| | Quality of College[13], [36], [95], [167], [175] | | |
| | Profession Interest[139] | Interpersonal Factors | Relationship with Administrative Staff[13], [27], [36], [135], [136], [137], [138] |
| Behavioral Factors | Engagement in Class[128], [132], [155], [156], [162], [168] | | |
| | Self-Confident [131], [139] | | |
| | Emotional Stability [131], [139] | | |
| | Self-Motivated[131], [135], [139], [159], [160], [161], [162] | Time Management Factors | Hours spent in Self-Study[140], [142], [155], [170], [174] |
| | Bad Habits [139] | | |
| | Hours spent in Social Sites[32], [154], [155], [156], [157], [158], [163], [164], [165], [166] | Health Factors | Disability [95] |
| | | | Health Issues[95], [128], [130], [131], [138], [173] |

## 2.6 MACHINE LEARNING TECHNIQUES FOR PREDICTING STUDENT ACADEMIC PERFORMANCE

ML models, both unsupervised and supervised, are specifically tailored for forecasting student academic success. Techniques such as Logistic Regression (LR), Decision Trees (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM), Naïve

Bayes (NB), Support Vector Machines with Linear Kernel (SVM-LK), XGBoost (XGB), Random Forest (RF), and K-Nearest Neighbors (KNN) are frequently employed to forecast student performance. Additionally, clustering techniques like K-Means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models (GMM), and Fuzzy C-Means (AM) are used to uncover hidden patterns in student data. These models assist educational institutions in optimizing academic support, enhancing resource allocation, and identifying at-risk students to implement timely interventions [5].

The following section reviews studies that have employed various machine learning techniques, presenting their methodologies, and key findings. These studies collectively highlight the merits and demerits of ML approaches in predicting student academic success. To consolidate the findings, a comparative summary table is provided at the end.

[4] developed a ML model that emphasized the necessity of preprocessing and algorithm fine-tuning techniques to address data inconsistent issues. The research employed three supervised learning algorithms, NNge, MLP, and J48, with a hybrid method employed to enhance classifier success. Experimental results demonstrated that J48 achieved the highest accuracy of 95.78% among the tested methods.

[6] proposed a model to forecast student academic achievement using SVM-LK, ID3, CART, and C4.5 classifiers in Weka. The findings proved that ID3 achieved 95.9% accuracy, SVM-LK achieved 96% accuracy, while C4.5 and CART both achieved 98.3% accuracy. C4.5 outperformed the others with higher precision (98.4%) and faster model-building time (0.03 seconds). CART and C4.5 had similar performance, with an incorrect classification rate of 1.7%. C4.5 demonstrated the best overall performance with reasonable model-building time.

[38] proposed an ML-based student prediction model using demographic, social, psychological, and economic factors. Eight ML classifiers, including LR, RF, SVM, XGB, SVM with a linear kernel, NB, KNN, and DT, were employed to forecast student performance.

[44] emphasized the use of Performance Factors Analysis (PFA) for adaptive

educational systems, known for its high prediction accuracy. To enhance PFA technically, the study introduced three ensemble learning- based approaches.

[53] investigated the use of logistic regression and SVM-LK to predict student performance based on feedback from 137 postgraduate students. The model, suited for binary outcomes, achieved a prediction accuracy of 72%. This study pinpoints the potential of online educational platforms to enhance teaching and learning processes and motivates educators to adopt such utilities for enhanced outcomes.

[54] proposed a hybrid machine learning approach combining Particle Swarm Optimization and Cuckoo Search to analyze features affecting education. This approach was applied to three datasets, identifying key factors and their relationships to individual performance. The optimal factors were used to predict academic outcomes using a regression neural network. The model outperformed with a high accuracy of 94%, proving the need of the hybrid optimization process in enhancing educational analysis and improving teaching methodologies.

[124] examined the uses of ML techniques in EDM, using data from virtual courses, e-learning logs, and student demographics to uncover meaningful patterns. The study employed ML techniques, including ANN, LR, and SVM, to forecast students' achievement depend on demographic characteristics and assignment scores. The results indicated that SVM outperformed the other models, achieving an accuracy of 87.3%.

[172] focused on identifying attributes that could predict student performance in online education during the COVID-19 pandemic, a period that saw a rapid shift to online learning due to the global crisis. The study employed five machine learning (ML) regressors to predict student performance, with support vector machine (SVM) achieving the highest accuracy of 87.5%. Additionally, a hybrid ensemble model that combined multiple classifiers was used to forecast student performance and produced outstanding results.

[176] studied student dropout in universities providing distance education and used the ML techniques to predict this issue. Several ML algorithms were tested, including NB, ANN, DT, and LR. The results indicated that NB outperformed the other algorithms, achieving an accuracy of 63%.

[177] introduced a prediction model for student academic performance using supervised ML algorithms, including SVM and LR. The study compared results from various experiments and demonstrated that the Sequential Minimal Optimization (SMO) algorithm within SVM outperformed LR, achieving an accuracy of 79%. The research highlighted impactful features such as teacher performance and student motivation, which can aid educational institutions in predicting and improving student outcomes while reducing dropout rates.

[178] highlighted the use of data analytics to predict student performance based on past academic experiences. The research proposed a new hybrid classification technique combining DT and Fuzzy Multi-Criteria Classification, achieving an accuracy of 82.28%. The model utilized criteria such as address, age, family size, activities, school, and previous grades.

[179] explored the application of ANN and comparing them with multivariate LR models. Three ANN models, Radial Basis Function (RBF), Probabilistic Neural Network (PNN), and Multilayer Perceptron (MLP) along with a SVM were employed. The comparison, based on the accuracy, revealed that ANNs achieved higher accuracy, ranging from 97.5% to 99.7%, outperforming LR.

[180] presented two prediction models to predict the students' performance. The study applied both SVM and KNN algorithms to predict students' grades. Results showed that SVM outperformed KNN, achieving a coefficient of 0.96 compared to KNN's 0.95. The study highlighted the usefulness of prediction models in enabling early interventions and task-specific student selection.

[181] developed an early-warning system to determine underperforming students and forecast learning outcomes in an LMS. Among the methods tested, the CART algorithm, enhanced by AdaBoost, emerged as the best classifier, achieving an accuracy of 78.45%. The system demonstrated its effectiveness in characterizing learning performance during a course, offering practical value for timely interventions.

[182] investigated the use of discussion forums as indicators of student performance, employing various data mining techniques to predict final performance. The study compared traditional classification with clustering-based and evaluated the effects of instance and attribute selection on accuracy. Among all approaches, the EM algorithm

achieved the highest accuracy of 90.2%.

[183] analyzed the performance of four ML algorithms to predict university student dropout. Results showed that RF, with 10 variables sampled at each split, was the most effective, achieving 91% accuracy and 87% sensitivity in a first validation sample.

[184] utilized EDM to develop a web-based system for predicting students' academic performance and identifying at-risk students using academic and demographic factors. Multiple ML algorithms, including SVM, RF, KNN, ANN, and LR, were applied, with the SVM model achieving a Mean Absolute Percentage Error (MAPE) of 6.34%. Results highlighted that academic factor, particularly midterm exam scores, had a more significant impact on performance than demographic factors.

[185] used ML to detect Autism Spectrum Disorder (ASD) and identify suitable teaching methods for affected children. The study combined ASD datasets and balanced them using SMOTE. The RF and XGB classifiers achieved 94% accuracy in ASD detection. The research also evaluated physical, verbal, and behavioral performance to develop educational approaches for children with ASD.

[186] proposed a combination prediction method for student performance based on ACO and ML algorithms. Three individual models were selected, with ACO determining the value of all the models. The combined model outperformed the individual models and other comparison methods, achieving a MSE of 0.0089, lower than that of the individual models and other methods. Additionally, the combination model demonstrated faster running speed.

[187] emphasized the critical importance of accurately predicting student academic performance to enhance educational outcomes and tailor learning methodologies. Their study focused on improving prediction precision by optimizing hyperparameters in ML models. Among various techniques compared, the Gradient Boosting Regression Tree (GBRT), optimized using the Grid Search (GS) method, achieved an accuracy of 93.6%. This study contributes to developing efficient and personalized educational strategies by enhancing prediction reliability.

[188] proposed a ML approach for predicting academic performance by incorporating education and socioeconomic data from Pennsylvania. The study used a range of

"black box" ML techniques to forecast academic performance. The neural network achieved the highest accuracy of 60% as compared to all the models.

**Table 2.11. Summary of the reviewed literature on ML Techniques**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [4] | 2019 | Developed a ML model that emphasized the necessity of preprocessing and algorithm fine-tuning techniques to address data inconsistent issues. | J48, NNge, MLP, Ensemble Method | J48 achieved the highest accuracy of 95.78%. | Did not extensively address scalability for larger datasets or complex feature interactions. |
| [6] | 2018 | Predicting student performance using ID3, C4.5, and CART classifiers in Weka. | ID3, SVM-LK, C4.5, CART | C4.5 achieved 97.3% accuracy, 97.45% precision, and the fastest model-building time (0.03 seconds). | No comparison with deep learning models. |
| [38] | 2024 | Developed a ML model using various factors to forecast students' performance. | LR, RF, SVM, XGB, SVM-linear kernel, NB, KNN, DT and ensemble DXK (DT + XGB + KNN). | Ensemble DXK achieved the highest accuracy of 97.83%. | Limited focus on feature importance and real-time application. |
| [44] | 2021 | Enhancing Performance Factors Analysis (PFA) with ensemble learning methods. | Random Forest, AdaBoost, XGBoost | XGBoost outperformed (93.5%) other models, improving predictive accuracy. | Limited exploration of technical improvements beyond pedagogical aspects. |
| [53] | 2022 | Predicted student performance based on feedback using a dataset of 137 postgraduate students. | Logistic regression and SVM-LK | LR achieved a prediction accuracy of 72%. | Limited to a small dataset; did not explore other machine learning models. |

| | | | | | |
|---|---|---|---|---|---|
| [54] | 2024 | Proposed a hybrid machine learning approach combining Particle Swarm Optimization and Cuckoo Search to analyze factors affecting education. | Regression Neural Network | Achieved 94% accuracy in predicting student performance by identifying optimal factors. | Limited to three datasets; could explore more diverse data sources |
| [124] | 2012 | Examined ML techniques in educational data mining using e-learning logs and demographics. | ANN, LR, SVM | SVM achieved the highest accuracy of 87.3%. | Did not address data imbalance or feature selection. |
| [176] | 2003 | Explored ML techniques to predict student dropout in distance education. | NB, ANN, DT, LR | NB achieved 63% accuracy and was implemented in a web-based tool. | Limited accuracy and focus on a single algorithm. |
| [177] | 2020 | Developed a supervised ML model for student academic performance prediction. | SVM and LR | SVM accuracy = 79%. | Lower accuracy compared to other models. |
| [178] | 2017 | Highlighted a hybrid classification technique for student performance prediction. | DT and Fuzzy Multi-Criteria Classification. | DT achieved 82.28% accuracy. | Limited integration of advanced ML methods. |
| [179] | 2016 | Compared ANN and LR models for predicting medical students' performance. | ANN (RBF, PNN, MLP) and SVM. | ANNs achieved 97.5% accuracy, outperforming LR. | Focused only on medical student data. |
| [180] | 2017 | Predicted final examination performance using ML algorithms. | SVM, KNN | SVM achieved a correlation coefficient of 0.96, outperforming KNN. | Limited algorithm comparison and feature analysis. |
| [181] | 2014 | Developed an early warning system for at-risk students. | CART with AdaBoost | Achieved 78.45% accuracy in identifying at-risk students. | Moderate accuracy compared to other studies. |
| [182] | 2013 | Investigated discussion forums as predictors of student performance. | DTNB, ADTree, LR, NB, clustering methods (EM, hierarchical). | EM algorithm achieved the highest accuracy of 90.2%. | Limited to discussion forums as predictors. |

| | | | | | |
|---|---|---|---|---|---|
| [183] | 2018 | Predicted university dropout using ML algorithms. | RF, NN, SVM, LR | RF achieved 91% accuracy and 87% sensitivity. | Focused on dropout without exploring other academic factors. |
| [184] | 2022 | Developed a system to predict student assessment and determine at-risk students. | RF, SVM, ANN, LR, and KNN | SVM achieved a MAPE of 6.34% and midterm scores were most influential. | Did not explore optimization techniques or advanced ensembles. |
| [189] | 2024 | Addressed imbalanced data and optimized ML algorithms for student performance prediction. | SMOTE, Bayesian Optimization, Grid Search, Random Search, RF, DT. | SMOTE with Bayesian Optimization and DT achieved 94.78% accuracy. | Focused on specific optimization techniques without ensemble comparison. |
| [185] | 2024 | Detected ASD and identified teaching methods using ML. | SMOTE, ensemble of RF and XGB. | Ensemble achieved 94% accuracy in ASD detection. | Limited focus on broader application in education. |
| [186] | 2024 | Proposed a combination prediction method for student performance. | ACO, DT, SVR, BP, NN | Combined model achieved MSE of 0.0089 and faster running speed. | Limited exploration of additional ML algorithms. |
| [187] | 2023 | Focused on improving the prediction of student academic performance to enhance educational outcomes. | Gradient Boosting Regression Tree | Achieved an accuracy of 93.6%, | Limited focus on other ML techniques and their comparative effectiveness. |
| [188] | 2023 | Predicted academic performance in Pennsylvania by integrating education. | DT, RF, LR, SVM, and NN. | Neural network achieved the highest accuracy of 60%, outperforming other models. | Accuracy improvements could be ma de with more diverse datasets. |

Researchers have also employed **optimization techniques** to enhance the performance of machine learning models, improving predictive accuracy and efficiency. Various studies have integrated these techniques to fine-tune parameters and optimize classification results. By systematically exploring and adjusting these parameters, these techniques ensure the model converges to the best solution, enabling better generalization to new data. Common techniques include Gradient Descent, which iteratively minimizes the loss function by adjusting parameters based on the gradient, and its variants like Stochastic Gradient Descent (SGD) and Mini-batch Gradient

Descent, which speed up computation by updating parameters using smaller batches or individual data points [190]. Adam and RMSProp [157] adjust learning rates adaptively, improving convergence in deep learning. L-BFGS is a quasi-Newton method that approximates the inverse Hessian matrix for optimization in smaller datasets [125]. Simulated and Vice-Simulated [54] are population-based methods inspired by natural processes, while Simulated Annealing explores the solution space probabilistically to avoid local minima [191]. Bayesian Optimization is used for hyperparameter tuning in complex models [192]. These methods enhance model accuracy, efficiency, and robustness.

[38] proposed a ML model to forecast student performance by analyzing demographic, social, psychological, and economic factors. Eight ML classifiers with nine feature selection techniques were employed to identify optimal factors for prediction. Among the models tested, the ensemble DXK achieved the outperformed accuracy of 97.83%, while the ACO-DT model, optimized using Ant Colony Optimization, further improved accuracy to 98.15%. This approach demonstrated superior performance in forecasting student outcomes and provided recommendations for future research directions.

[61] presented a recommendation system for student courses using ACO. The study demonstrated the effectiveness of ACO in predicting students' final grades upon completing university courses. ACO was applied by representing the problem as a graph, where each node symbolized a decision in the domain. The method showcased promising results, highlighting ACO's potential in course recommendation systems.

[62] proposed a personalized learning path planning method using ACO to address challenges in online education. The study utilized dynamic time regularization and K-means cluster to identify students' learning types, followed by ACO to generate optimal paths. The method demonstrated superior performance, achieving a classification accuracy of 96.6%, a course completion rate of 90%, and an average video learning time of 80 minutes, outperforming GA-based approaches by 20%. This approach significantly enhances academic performance and online learning outcomes.

[193] developed two hybrid wrapper-feature-selection methods combining ACO and Hybrid Rice Optimization (HRO) to address challenges in high-dimensional data

analysis, particularly for biomedical applications. The first hybrid embeds ACO as an evolutionary operator within HRO, while the second uses independent subpopulations that share local search results. A heuristic factor assignment strategy enhances ACO's global search for optimal features. Evaluated on fourteen biomedical datasets, the methods demonstrated improved accuracy and computational efficiency compared to other advanced feature selection algorithms.

[58] proposed a novel method for developing adaptive IIR and FIR filters using the ABC algorithm. This approach addressed challenges associated with the multi-modal error surfaces of adaptive IIR filters, where gradient-based methods often fail due to local minima, leading to instability and suboptimal convergence. By utilizing the global optimization capabilities of the ABC algorithm, the proposed method achieved enhanced stability and optimal performance. Simulations for noise cancellation demonstrated that this method outperformed traditional gradient-based and evolutionary approaches.

[59] proposed a Bee Colony Optimization Metaheuristic (BCO) approach. The study introduced a heuristic algorithm for the Ride-Matching Problem as an illustrative example, highlighting the capabilities of BCO. Results demonstrated its effectiveness in addressing complex optimization problems under both deterministic and uncertain conditions, showcasing its potential for diverse real-world applications.

[194] presented an enhanced ABC algorithm tailored for constrained optimization problems using Deb's rule. Results demonstrated that the improved ABC algorithm achieved comparable solution quality while exhibiting faster convergence compared to the Particle Swarm algorithm.

The development of **web-based predictive interfaces** has gained significant attention in educational research, enabling institutions to leverage advanced machine learning (ML) and deep learning (DL) algorithms for academic assessment prediction.

Additionally, researchers have examined different tools and programming languages for implementing web-based interfaces that effectively deploy ML models. Several studies have utilized frameworks like Django, Flask, and Node.js to integrate predictive models into interactive platforms, facilitating real-time student performance assessment and decision-making support. By implementing web interfaces, these systems provide real-time predictions of students' academic performance and generate actionable insights [47], [128], [195].

**Table 2.12. Summary of the reviewed literature on ACO technique**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [38] | 2024 | Proposed a ML model to forecast student's success by analyzing demographic, social, psychological, and economic factors. | ML classifiers (DT, XGB, KNN) and feature selection techniques; ACO-DT for optimization. | DXK achieved 97.83% accuracy; ACO-DT improved to 98.15% accuracy. | Need to validate the model across diverse datasets and institutions. |
| [61] | 2010 | Presented a recommendation system for student courses using ACO. | ACO represented as a graph for course recommendations. | Highlighted ACO's potential in improving course recommendation systems by predicting final grades with promising accuracy. | Integration of additional contextual factors in course recommendations. |
| [62] | 2024 | Proposed a personalized learning path planning method using ACO to address challenges in online education. | Dynamic time regularization, K-means clustering for classification, ACO for optimal path generation. | Achieved 96.6% classification accuracy, 90% course completion rate, and 80 min video learning time, outperforming GA-based methods by 20%. | Validation in diverse educational environments and student populations. |
| [193] | 2023 | Proposed two hybrid wrapper feature selection methods combining ACO and Hybrid Rice Optimization (HRO) for high-dimensional biomedical data. | ACO embedded within HRO; subpopulation-based hybridization; heuristic factor assignment for ACO. | Demonstrated improved accuracy of 84.3% by C-IBACO and computational efficiency across fourteen biomedical datasets compared to other advanced algorithms. | Application to non-biomedical datasets and real-time systems. |

**Table 2.13. Summary of the reviewed literature on ABC technique**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [58] | 2011 | Developed a novel approach for developing adaptive IIR and FIR filters using the ABC algorithm. | ABC | Enhanced stability and convergence; outperformed gradient-based and evolutionary approaches in noise cancellation simulations. | Lack of exploration in broader domains beyond noise cancellation. |
| [59] | 2003 | Proposed a Bee Colony Optimization Metaheuristic (BCO) approach for solving complex optimization problems. | Bee Colony Optimization | Effectively solved deterministic and uncertain optimization problems; demonstrated versatility in real-world applications. | Limited exploration of scalability and adaptability in larger datasets or complex systems. |
| [194] | 2003 | Presented an enhanced ABC algorithm for constrained optimization. | Modified Artificial Bee Colony (ABC) with Deb's rule, CMC. | CMC achieved 86.8% accuracy comparable solution quality with faster convergence. | Limited testing on diverse real-world engineering problems. |

[160] proposed a method for predicting student behavior in web-based e-learning systems by analyzing feedback and log data. The study focused on predicting whether students would continue their engagement or leave the platform. A classifier based on polynomial regression and stochastic gradient descent was used to process user data in a single iteration, reflecting dynamic behavior changes. The approach helped improve student retention by offering personalized interventions.

[184] developed a web-based system for at-risk students using demographic and academic factors. The study employed several ML algorithms to predict total course scores at early stages and achieved a MAPE of 6.34%. Results emphasized that academic factors, particularly midterm exam scores, had a highest impact on student's success compared to demographic factors. The web-based system, hosted on an online server, offers tutors actionable insights to better support students' academic success.

[194] developed a method to predict students' final grades using data from a web-based

education system. The study evaluated various pattern classifiers and found that combining them improved performance. The use of a genetic algorithm (GA) to weight features further enhanced prediction accuracy by 10-12%. This approach helps identify at-risk students early, enabling timely intervention, especially in large classes.

[196] explored AI-powered web-based educational tools designed to improve learning and teaching experiences. The study highlighted applications such as predictive analytics, automated grading systems, and AI-enhanced scaffolding for scientific writing. It emphasized the role of web-based platforms in supporting educators with real- time insights and personalized learning approaches while addressing ethical challenges and risks.

[197] proposed an adaptive learning system to personalize education by tailoring learning materials and assessment strategies to students' characteristics. The study utilized VAK learning styles to classify learners based on their learning activity clicks, integrating machine learning algorithms with Python and semantic association to map learning activities to VAK styles.

[198] aimed to explore students' perceptions of web-based versus paper-based homework and identify differences based on homework performance and GPA. The results revealed no significant difference in GPA scores. Additionally, the study concludes with some tentative recommendations for future practices.

[199] proposed an educational data mining approach to predict poor-performing students and enhance academic outcomes. The study conducted a comparative analysis of prediction models, with the hybrid random forest model (Hybrid RF) demonstrating the best classification performance. The developed system, utilizing this technique, provides early predictions of student performance, enabling timely interventions by educational stakeholders. Experimental results confirmed the system's effectiveness in improving academic performance prediction and facilitating targeted interventions.

[200] conducted a study focusing on predicting student dropout rates by analyzing web-based course activity. The study identified that changes in student activity on course websites could help in early detection of at-risk students, allowing educators to intervene before actual dropouts occur. The research found that 66% of students

flagged as at-risk based on their behavior patterns did not complete their courses or degrees. This analysis helps instructors identify students with unusual activity patterns and enables institutions to implement timely interventions.

**Table 2.14. Summary of the reviewed literature on web-based prediction interface**

| Ref. | Year | Study Description | Method/ Algorithm Used | Key Findings | Gap |
|---|---|---|---|---|---|
| [160] | 2016 | Proposed a method for predicting student behavior in web-based e-learning systems. | Python and Pycharm | Helped improve student retention by predicting whether students would continue engagement or leave the platform. | Focus on prediction, but lacks in-depth intervention strategies. |
| [184] | 2022 | Developed a web-based system for at-risk students using demographic and academic factors | Python and Flask | Achieved a MAPE of 6.34% and midterm exam scores, had a greater impact on performance. The system offers actionable insights for tutors. | No explicit gap identified. |
| [194] | 2003 | Developed a method to predict students' final grades using web-based system data. | Java and Springboot | Combining classifiers improved performance. GA enhanced prediction accuracy by 10-12%. | Limited to web-based system data, not considering external factors. |
| [196] | 2024 | Explored AI-powered web-based educational tools to enhance teaching and learning experiences. | JavaScript and Express.js | Highlighted the role of web-based platforms in providing real-time insights and personalized learning while addressing ethical challenges. | Ethical challenges and risks remain underexplored. |
| [197] | 2024 | Proposed an adaptive learning system to personalize education by tailoring learning materials and assessments. | Python and Pycharm | The model was validated with the OULAD, demonstrating effectiveness in enhancing adaptive learning experiences. | Limited scope of datasets used. |
| [198] | 2007 | Aimed to explore students' perceptions of paper-based homework and web-based homework. | PHP and Laravel | No significant difference in GPA. | Lack of in-depth analysis on the reasons behind student preferences. |

| | | | | | |
|---|---|---|---|---|---|
| [199] | 2020 | Proposed an educational data mining approach to predict poor-performing students and enhance academic outcomes. | Python and Pycharm | Hybrid RF demonstrated the best classification performance. The system enables early prediction of student performance and facilitates timely interventions. | Need for more robust feature selection techniques. |
| [200] | 2017 | Conducted a study on predicting student dropout rates. | Python and Pycharm | Identified that changes in student activity could help in early detection of at-risk students. | Need for more nuanced behavior data. |

## 2.7 SUMMARY

This chapter presents a comprehensive review of existing studies related to the forecasting of student academic performance using ML techniques. The goal of the review is to analyze various algorithms, methodologies, and approaches adopted in the literature to address challenges in academic performance prediction. The studies explore diverse datasets to predict performance metrics such as grades, scores, and pass/fail status.

A wide range of machine learning techniques has been utilized, such as KNN, DT, RF, SVM, NB, LR, GB, and ANN. Some studies also incorporate advanced methodologies, such as hybrid models and optimization techniques, to enhance predictive accuracy.

Throughout the chapter, the focus is placed on identifying gaps in the literature, such as limited dataset diversity, lack of real-time applications, and minimal emphasis on scalability. Additionally, the review in Chapter 2 emphasizes the relevance of using ML models for academic performance prediction and aligns these findings with the research objectives outlined in Chapter 1.

The insights gained in this chapter guide the development of the methodology presented in Chapter 3, where the proposed approach for student performance prediction is detailed. The outcomes of the implementation are presented and compared with existing methods in Chapter 4. In conclusion, Chapter 5 discusses the implications, conclusions, and potential future directions for this research.

## Chapter 3

# RESEARCH METHODOLOGY

Educational research focuses on academic performance prediction as an essential issue because it helps identify struggling students so they can receive immediate assistance. A systematic, multi-stage methodology must be employed to obtain credible findings during academic performance forecasts. The methodology chapter presents every aspect related to building the prediction model from scratch.

The research begins with data collection and preparation, which involves cleaning, transforming, and preparing raw student response data for analysis. The selection of features through this process helps determine academic success factors and also enhances model accuracy by reducing dataset dimensions. The research utilizes complex selection methods that choose optimal features, as outlined in this chapter.

The next phases involve both model development and optimization. This research implements eight ML classifiers and two optimization techniques employed for top-model optimization: Ant Colony Optimisation (ACO) and Artificial Bee Colony (ABC). These methods boost predictive accuracy while decreasing the number of errors that occur.

This chapter also explains how a user-friendly web interface implements the predictive model for accessible system usage. Through the interface, stakeholders can view results containing recommendations while conducting both bulk and individual student forecasting.

## 3.1 BACKGROUND STUDY

The identification of failing students, together with improved outcomes for institutions, makes academic performance prediction a leading research topic in education. Heritage combined with behavioural patterns along with interpersonal and familial elements form only a small subset of numerous variables that affect a student's educational outcome. A thorough evaluation of these attributes helps students and educators gain a better understanding of academic traits and potential.

Machine learning algorithms have gained popularity in classroom performance prediction in recent years. Large datasets, combined with the capability of ML models to find intricate relationships between characteristics, make them superior to traditional forecasting approaches. Many algorithms have proven successful in predicting academic achievement among students. The models involved include NB, SVM, DT, XGB, and RF. The performance quality of these models increases substantially when optimization methods ABC and ACO are applied, leading to improved prediction accuracy.

Additionally, predictive models have expanded their real-time applications through web-based platforms. Organizations operating higher education institutions leverage these platforms to gather student feedback while developing personalized academic improvement plans. These technological systems allow administrators, together with teachers, to provide early assistance for underperforming students, thus supporting recruitment efforts on campus and educational development programs.

This research develops an extensive predictive ML model that uses multiple parameter dimensions to forecast performance results. The research aims to build a predictive system with precise results through the combination of optimization methods, ML classifiers, and feature selection techniques. The model functions as a practical tool because its developers made it available through a web-based platform that enables effective usage in educational settings.

## 3.2 PROBLEM FORMULATION

Campus recruiting success together with career development for students depends on accurate forecasting of their educational progress. Various factors, including poor time management, family, socio-economic and financial troubles, and insufficient personalized guidance, cause students to face unemployment by being unable to achieve a satisfactory recruitment benchmark. Traditional approaches to academic performance prediction yield unsatisfactory results due to their reliance on small datasets and their inability to capture the multidimensional nature of students and other important factors.

An efficient predictive model becomes essential because it must identify academic difficulties, detect underperforming students, and offer helpful recommendations. The model needs to achieve higher accuracy by integrating demographic, socio-economic, and behavioural factors, along with advanced ML approaches.

The inability of traditional techniques to incorporate optimization methods further limits their application in real-world educational settings. Previous researchers could have applied advanced optimization strategies, including ACO and ABC techniques, to improve the functionality of existing ML models, but these were overlooked.

This research presents an academic success prediction model that applies optimized machine learning as an approach to resolve this educational issue. The model focuses on university or college students' academic assessments to determine whether students will achieve results above 60%, a minimum criterion for recruiters in campus placements. The model's accuracy increases when optimization methods boost its prediction performance, along with feature selection methods such as wrapper and embedding strategies that identify optimal academic achievement factors.

Additionally, this study also builds a web-based platform that executes the optimized prediction model. The platform enables real-time evaluation of student responses, allowing institutions to provide personalized recommendations to their students. This integration ensures both the practical feasibility and scalability of the solution for educational purposes.

By combining feature selection methods and optimized machine learning classifiers, this study aims to overcome the limitations of traditional methods and contribute to improving academic outcomes through an accurate and practical solution.

## 3.3 TECHNIQUES USED

The following section highlights the various methods implemented to achieve the research objectives of this study.

### 3.3.1 Data Imbalance Technique (SMOTE)

Through adding synthetic features SMOTE enhances the model's performance on under-represented samples in the data. SMOTE uses the numeric differentiation between current minority class instances and each of their nearest neighbours to create simulated data points. The following formula shows the procedure for making synthetic data:

$$Xnew = Xi + \lambda \cdot (\text{Xneighbor} - \text{Xi}) \tag{3.1}$$

The current class instance (Xi) feature-vector combines with the feature-vector from its nearest neighbour (Xneighbor) through a random interpolation factor (λ) that falls between 0 and 1. Implementing SMOTE ensures a balanced dataset, which in turn improves classifier performance.

### 3.3.2 Feature Selection Techniques

Feature selection was performed to identify the most influential factors impacting student academic performance. The following nine techniques were used,

#### A. Chi- Square Test ($x^2$)

The Chi-Square test is a statistical method used to determine the relationship between a feature and the target variable by assessing their independence. It calculates the degree of association between categorical variables using the formula:

$$x^2 = \frac{\sum(Observedi - Expectedi)^2}{Expectedi} \tag{3.2}$$

where, Observedi is the observed frequency and Expectedi is the expected frequency of the $i^{th}$ category with no association between the feature and the target. A higher $x^2$ value indicates a stronger correlation between the target feature and other feature.

#### B. Pearson Correlation Coefficient (rXY)

This statistical metric is employed to ascertain the direction and magnitude of the linear relationship between features, X and Y. It calculates using the following formula:

$$rXY = \frac{\sum\left((Xi - Xmean).(Yi - Ymean)\right)}{\sqrt{\sum(Xi - Xmean)^2 . \sum(Yi - Ymean)^2}} \tag{3.3}$$

where, the sample points are Xi and Yi, while the variable means are Xmean and Ymean. The covariance between X and Y is computed by the numerator, and it is normalised by the product of their standard deviations in the denominator. The resultant number falls between -1 and 1, with values around 0 indicating little to no linear correlation and values closer to 1 or -1 indicating a significant positive or negative linear association.

### C. Variance Threshold ($Var(X)$)

A straightforward method called the Variance Threshold eliminates features with low variance since it is assumed that features with low variability are less useful for classification tasks. The method calculates the variance of each feature and removes those with variance below a specified threshold. The formula for variance is:

$$Var(X) = \frac{1}{N} \sum N_{i=1} (Xi - Xmean)^2 \tag{3.4}$$

where, N is the number of samples, Xi is the $i^{th}$ sample value of the feature, and Xmean is the mean of the feature values. A feature is selected if its variance is above a pre-defined threshold (θ) value (Var(X)> θ).

### D. Feature Importance ($I(f)$)

Feature importance quantifies how much a feature contributes to a model's predictions. It is used to determine the relevance of each feature in the decision-making process. For a feature f, the importance I(f) using Gini Impurity is computed as Eq. (3.5),

$$I(f) = \sum_{t \in Tf} \frac{Nt}{N} \left(Gini(t) - Gini(tL) - Gini(tR)\right) \tag{3.5}$$

where, Tf is the set of all nodes where feature f is used for splitting, total number of samples (N), The number of node samples t is Nt, Gini(t) is the Gini impurity of node t, tL and tR are the left and right nodes, and the Gini impurities of the right and left nodes are Gini(tR) and Gini(tL), respectively.

### E. RFE ($ri$)

RFE is a technique that repeatedly removes the least significant variables based on a given model's significance ranking, with the goal of improving the model's success rate. The mathematical representation of this technique can be summarized as Eq. (3.6),

$$ri = \sum_{j \in Xi} |\, wj \,| \tag{3.6}$$

where, ri is the rank of the feature set Xi, Xi is the set of features at the $i^{th}$ iteration, and wj is the weight or importance score of features Xj. The goal is to iteratively eliminate the feature xk with the smallest |wk|, which can be represented by,

$$k = arg\ min_{j \in Xi} |\, wj \,| \tag{3.7}$$

After identifying the feature xk, it is removed from the feature set for the next iteration,

$$X_{i+1} = \frac{Xi}{\{xk\}} \tag{3.8}$$

Continue this procedure until the required number of variables m is reached, $|\, X \,| = m$. These equations iteratively eliminate the feature with the smallest importance score to progressively refine the feature set.

### F. Lasso ($\boldsymbol{\beta}$^)

Lasso regression is a kind of linear regression that adds regularisation and variable selection to improve the statistical model's interpretability and prediction accuracy. The mathematical representation of this technique can be summarized as Eq. (3.9),

$$\beta^{\wedge} = \arg min_{\beta} \{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} \beta j xij )^2 + \lambda \sum_{j=1}^{p} |\, \beta j \,| \} \tag{3.9}$$

where, $\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} \beta j x i j)^2$ represents the sum of squared errors, $\lambda \sum_{j=1}^{p} \mid \beta j \mid$ is the L1 regularization term, where λ controls the strength of the penalty on the coefficients βj, and the objective is to identify the coefficients $\beta^{\wedge}$ that minimize this equation.

### G. Ridge ($J(\beta)$)

Ridge Regression adds an L2 penalty term to the cost function in order to address overfitting and multicollinearity in regression models. This penalty discourages large coefficients, ensuring a more stable and generalizable model. The ridge coefficient estimates are obtained by minimizing the following objective function,

$$J(\beta) = \sum_{i=1}^{n}\left(y_i - \beta 0 - \sum_{j=1}^{p} \beta j x i j\right)^2 + \lambda \sum_{j=1}^{p} \beta j^2 \tag{3.10}$$

where, J(β) is the cost function, observed value of the dependent feature is yi, β0 and βj are the intercept and coefficients, respectively. The xij denotes the $j^{th}$ feature of the $i^{th}$ data point and the parameter controlling the value of the penalty is $\lambda$ .

By adding the penalty term $\lambda \sum_{j=1}^{p} \beta j^2$, Ridge Regression reduces the magnitude of the coefficients, helping to prevent overfitting while maintaining model interpretability. A higher λ value increases the penalty, shrinking coefficients closer to zero. However, unlike Lasso regression, Ridge does not reduce coefficients to exactly zero, making it suitable for models where all features are relevant.

### H. Random Forest ($\boldsymbol{Imp(f)}$)

Random Forest can use both the Gini index and entropy to measure feature importance. Entropy evaluates the impurity or disorder in the dataset and is calculated at each node split. The feature importance is derived from the reduction in entropy when a feature is used to split the data. The mathematical representation using entropy is provided in Eq. (3.11).

$$Imp(f) = \sum_{t \in Tf} \frac{Nt}{N} \left(Entropy(t) - Entropy(tL) - Entropy(tR)\right) \tag{3.11}$$

where, Tf is the set of all nodes where feature f is used for splitting, N is the total number of samples, Nt is the number of samples at node t, Entropy(t) is the entropy at node t, tR and tL are the right and left nodes of t, and Entropy (tR) and Entropy(tL) are the entropy of the right and left nodes, respectively.

The Entropy(t) is obtained by,

$$Entropy(t) = -\sum_{i=1}^{C} pilog_2(pi) \tag{3.12}$$

where, C is the total classes and pi is the percentage of samples in class i at node t. A feature f is considered important if it results in a significant reduction in entropy when used for splitting.

### I. XGBoost

XGBoost, a powerful gradient-boosting framework, evaluates feature importance using metrics like Gain, Cover, and Frequency. For feature selection, Gain is commonly used as it quantifies the improvement in the model's accuracy when a feature is used for splitting. A higher Gain value indicates that the feature contributes significantly to reducing the error. The representation of Gain is represented by Eq. (3.13).

$$Gain = Gleft + Gright - Gnode \tag{3.13}$$

where, Gnode is the impurity measure (e.g., log loss) at the current node, Gleft and Gright are the impurity measures after splitting the node. The features with the highest Gain values are prioritized for selection.

### 3.3.3 Classification Techniques

The use of classification techniques is essential for developing predictive models that divide academic achievement into discrete groups, such as passing or failing. They help identify patterns in the data that can provide valuable insights for targeted interventions. These methods utilize input features and a labeled dataset to learn patterns that differentiate between the defined classes. The following eight techniques are used in this research.

### A. Decision Tree

In the tree-structured model called as the DT classifier, every node stands for a feature, every branch for a decision rule, and every leaf for a result. For student performance prediction, the DT uses training data to create splits based on features that minimize impurity. The Gini Index for a node is computed as,

$$Gini\ (t) = 1 - \sum_{i=1}^{c} pi^2 \tag{3.14}$$

where, the probability of class is pi, and the number of classes is c. The feature that maximally reduces impurity at each node is chosen for the split, making DT highly interpretable and effective for predicting whether students score above or below the threshold.

### B. Naïve Bayes

NB is a probabilistic model based on Bayes' Theorem, which assumes independence among features. For predicting academic performance, NB determines the possibility of every class (e.g., passing or failing) given the features. The posterior probability of a class y given the features X=(x1,x2,...,xn) is,

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{3.15}$$

Here, the likelihood is P(X|y), the prior is P(y), and the evidence is P(X).

### C. K-Nearest Neighbour

In order to predict the class of a newly-inserted data point, KNN looks to the training data to find out which k closest neighbours the data point has. Using the student's feature vector and the dataset's average similarity (as measured by the Euclidean distance), KNN predicts the student's academic success.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}}(xi - yi)^2 \tag{3.16}$$

where xi and yi are the feature values of two samples.

## D. Support Vector Machine

Support vector machines (SVMs) construct a hyperplane that maximizes the margin between fail and pass classes in order to forecast students' academic achievement. The optimization problem is represented as,

$$min\ \frac{1}{2}||w||2\ subject\ to\ yi\ (w \cdot xi + b) \geq\ 1, \forall I \tag{3.17}$$

where, the weight vector is w, the bias is b, the feature vector is xi, and the class label (-1 for failing and +1 for passing) is yi. This method works well with datasets that have a complicated relationship between attributes and outcomes.

## E. Support Vector Machine-Linear Kernal

A following linear kernel-based support vector machine (SVM) extends SVM to handle non-linear decisions by using kernel functions to transform input to a higher-dimensional space. It is common practice to use the Radial Basis Function when classifying data.

$$K(x, x') = \exp\left(-\gamma\big||x - x')\big|\right| \tag{3.18}$$

where, x′ and x are the data-points, and γ controls the kernel's flexibility.

## F. Logistic Regression

The following sigmoid function determines the probability of either passing or failing by analyzing the inputs X and weights w through a LR model,

$$P(y = 1|X) = \frac{1}{1+e^{-(w.X+b)}} \tag{3.19}$$

The vector w represents weights while X designates features along with Bias term b. A training cost function functions as a reliable forecasting method for student performance analysis through minimizing prediction gaps with actual results.

### G. XGBoost

XGB functions as a combination of ensemble methods built on decision trees. XGB builds multiple trees through iterative processes, using the following log-loss function to predict academic success,

$$L = -\sum_{i=1}^{n}[yilog(\hat{y}i) + (1 - yi)log(1 - \hat{y}i)] \quad (3.20)$$

where yi is the actual class, and $\hat{y}i$ is the predicted probability. The method of model development through XGB proves effective as it handles incomplete data while optimizing class prediction accuracy.

### H. Random Forest

Model generalizability and performance stability increase through the deployment of bagging and related techniques into RF models. The categorization scheme is built from the majority votes provided by the trees. The model calculates the actual achievement forecast through,

$$\hat{y} = mode\ (\{T1(X), T2(X), \ldots, Tk(X)\}) \quad (3.21)$$

The prediction of the $i^{th}$ tree appears as Ti(X) in the mathematical expression. The RF model provides excellent analytical capabilities for complex student performance data by reducing overfitting problems.

## 3.3.4 Model Optimization Techniques

### A. Artificial Bee Colony

Artificial Bee Colony (ABC) is a swarm intelligence-based technique that optimizes model hyperparameters by simulating the exploration and exploitation of nectar sources. The optimization process can be summarized mathematically as follows,

For Position Update,

$$\vartheta ij = x_{ij} + \emptyset_{ij}(x_{ij} - x_{kj}) \tag{3.22}$$

where, $\vartheta ij$ is the new position of the solution, $x_{ij}$ is the current position of the solution, $x_{kj}$ is the position of a randomly selected solution k, and $\emptyset_{ij}$ is the random number in the range [-1,1].

For Fitness Evaluation,

The fitness $f_i$ of a solution is calculated as,

$$f_i = \{\frac{1}{1+f(xi)}, if\ f(xi)\ \geq 0\ and\ 1 + |f(xi)|, if\ f(xi) < 0\} \tag{3.23}$$

where f(xi) is the objective function value for solution xi.

For Probability of Selection,

$$Pi = \frac{fi}{\sum_{j=1}^{N} fi} \tag{3.24}$$

where, the probability of selecting solution i is Pi and N is the total number of solutions. ABC optimizes the model by refining hyperparameter values through the iterative collaboration of bees, leading to better classification performance.

### B. Ant Colony Optimization

Ant Colony Optimization (ACO) is a optimization technique used to improve the hyperparameters of ML models. In the context of model optimization, ACO finds the optimal set of hyperparameters to improve classification performance. The optimization process is mathematically represented as follows,

For Pheromone Update Rule,

$$\tau ij(t+1) = (1-\rho) \cdot\ \tau ij(t) + \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \tag{3.25}$$

where, the pheromone level is $\tau ij(t)$ at time t, evaporation rate is $\rho$, m is the number of ants, and $\Delta\tau_{ij}^{k}$ is the pheromone deposited by ant k on the edge between i and j.

For Probability of Path Selection,

$$P_{ij}^{k} = \frac{\tau ij(t)^{\alpha} \cdot \eta ij^{\beta}}{\sum_{l \epsilon J_i^k} \tau il(t)^{\alpha} \cdot \eta il^{\beta}} \tag{3.26}$$

where, $\eta ij$ is the heuristic value, $\alpha$ is the weight of pheromone importance, $\beta$ is the weight of heuristic importance, and $J_i^k$ is the set of feasible moves for ant k from node i. ACO optimizes the machine learning model by iteratively refining hyperparameter values based on the performance of each ant's solution. This leads to an improved and robust model.

### 3.3.5 Evaluation Metrices

Evaluation metrics are critical for assessing the performance of classification models, offering insights into their forecasting capabilities. The following metrics were used in this research,

#### A. Accuracy

The percentage of correctly identified occurrences relative to all records in the dataset is known as accuracy. The formula is depicted as Eq. (3.27),

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.27}$$

where, True Positive is TP, True Negative is TN, False Positive is FP, and False Negative is FN. A higher value represents that the model correctly classifies most instances.

#### B. R-Score ($R^2$)

A model's explanatory power regarding data variation can be measured through the R-Score which is known as the coefficient of determination. The R-Score acts as the perfect tool for regression-based evaluation in classification problems. The formula is depicted as Eq. (3.28),

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(yi-\widehat{yi})^2}{\sum_{i=1}^{N}(yi-\overline{yi})^2} \tag{3.28}$$

where, the actual value is yi, the predicted value is $\hat{y}\imath$, and the actual values mean is $\bar{y}\imath$. The model is more accurately fit when the $R^2$ score is close to 1.

### C. F-score (F1)

The F-score is a balanced metric that accounts for both false positives and false negatives. The formula is depicted as Eq. (3.29),

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.29}$$

This metric is useful for imbalanced datasets. A higher F-score indicates better performance in differentiating classes.

### D. Recall

The recall measures how well the model can identify positive instances. It is defined as,

$$Recall = \frac{TP}{TP+FN} \tag{3.30}$$

A high recall shows that the model can accurately identify positive cases, reducing false negatives.

### E. Precision

Precision is the percentage of accurately predicted positive instances out of all instances predicted as positive. The formula is,

$$Precision = \frac{TP}{TP+FP} \tag{3.31}$$

The model's ability to reduce false positives and maximize true positive predictions is reflected in its high precision.

**F. Confusion Matrix**

A confusion matrix is a performance summary that shows the correct and incorrect predictions made by a classification model. It is defined as,

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{3.32}$$

where, TP is the true positives, FP is the false positives, FN is the false negatives, and TN is the true negatives.

**G. ROC Curve**

The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various levels. It is defined as,

$$TPR = \frac{TP}{TP+FN} \tag{3.33}$$

and,

$$FPR = \frac{FP}{FP+TN} \tag{3.34}$$

A curve closer to the top-left corner indicates better performance in identifying positive and negative classes.

**H. AUC (Area Under the Curve)**

AUC measures the entire area under the ROC curve, providing a single value to summarize model performance. It is defined as,

$$AUC = \int_0^1 TPR(FPR)d(FPR) \tag{3.35}$$

A higher AUC value (closer to 1) indicates that the model can effectively distinguish between positive and negative classes.

### 3.3.6 Web-Based Prediction System

The predictive model for students' academic achievement was implemented using a web-based prediction system. This Django-based system offers real-time predictions using either individual student data input or batch processing through Excel files. The system depends on the optimized machine learning model to predict whether students will achieve scores above or below 60%. Users can easily access results with recommendations through a dedicated section, where they can click the "Predict" button to generate results. Additionally, a "Generate Report" button is available in the system, allowing users to create comprehensive PDF reports that can be analyzed for further evaluation.

## 3.4 RESEARCH DESIGN

The standard research method used to forecast academic success in students involves thorough organizational procedures for developing stable prediction models. This method breaks down data processing for academic success forecasting into multiple stages to achieve accurate results. The research merges multiple datasets through sophisticated ML approaches to enhance the effectiveness of prediction models in evaluating students and addressing their complexities. Figure 3.1 depicts the entire methodology employed to forecast student performance at a high organizational level.

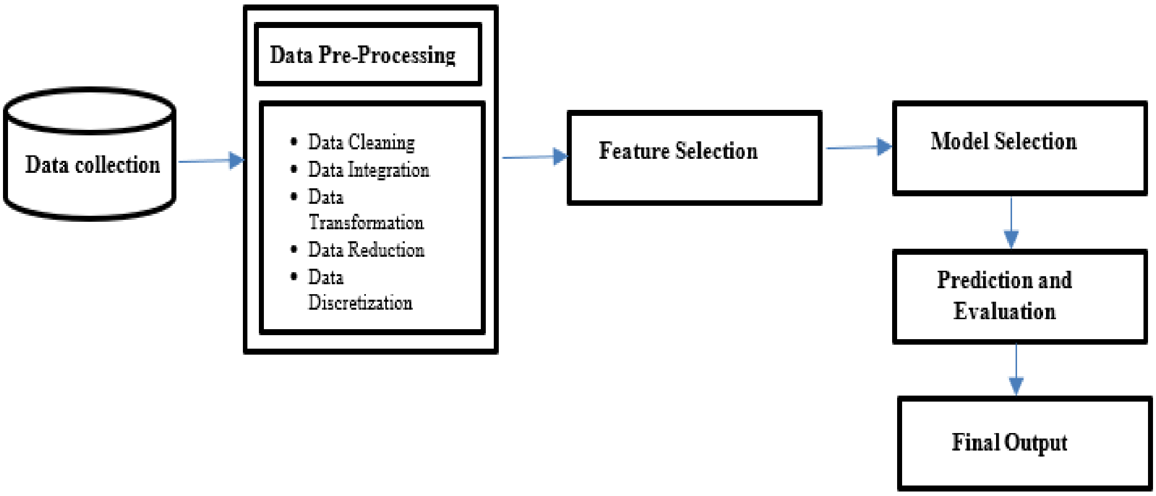


**Figure 3.1. General steps in student academic performance prediction**

**Step 1: Data Collection,** as the first step, requires gathering student information by using questionnaires and institutional records as well as surveys. This array of features includes demographic, socio-economic, behavioral, and academic features that are crucial for predicting student performance.

**Step 2:** Several **Data Preprocessing** operations consisting of filtering, integrating, transforming and reducing and discretizing the data need to be applied to unprocessed data before analysis can begin. The data cleanup process removes incorrect information as well as duplicates and missing values, while the integration step combines data from separate sources. Also, during this phase, the categorization of discrete attributes, numerical scale adjustment, and formatting standardization take place. The reduction of data occurs through dimensionality reduction combined with feature selection techniques that remove unnecessary or less important information. The data needs preparation for model training through variable categorization processes when necessary.

**Step 3: Feature Selection** represents the next stage in predicting a student's academic achievement because it helps identify the most important features. Through this approach, researchers attain better model performance, reduced computing complexity, and eliminate useless features. Feature selection was conducted on the training set only, and the chosen features were used on the test set for evaluating the model.

**Step 4:** The next step is **Model Selection**, when several machine learning techniques are assessed to see which model best categorizes student performance.

**Step 5:** The second-to-last step is **Prediction and Evaluation**, when the trained model is evaluated on new data. Evaluation metrics including ROC curve, precision, accuracy, recall, R-score, and F1-score are employed to assess the model's efficiency in predicting academic performance.

**Step 6:** The last step, **Final Output** involves deploying the prediction system, which generates actionable outputs, such as predicting whether students will meet a specific performance threshold, providing personalized improvement recommendations, and offering a user-friendly web-based interface for real-time predictions and decision-making.

### 3.4.1 Dataset Used

The dataset used in this research includes records of second-year undergraduate Computer Science students from St. Aloysius' College (SAC). This primary dataset contains 1369 student records and 70 features, capturing a diverse range of attributes, including demographic, academic, socio-economic, and behavioral factors that affect the performance. Among these, the 28 red-highlighted features (Table 3.1) were novel features that had not been considered by any researchers previously, while the remaining 42 features were adopted from state-of-the-art research works.

The original dataset (Table 3.1), named student_data.csv includes features such as age, gender, attendance, previous academic performance, socio-economic conditions, learning behavior, and parental education levels. Feature selection approaches were used to find the most pertinent attributes in order to improve the model's accuracy and efficiency. As a result, the dataset was reduced to 21 significant features (excluding the target variable) (Table 3.2), which were used for model training and evaluation. The final refined dataset is named stu_dataset_final.csv, representing the most critical predictors of academic performance. The feature selection process eliminated irrelevant and redundant features, thereby reducing computational complexity and improving model performance.

### A. Additional Datasets for Web-Based Prediction Implementation

In addition to the primary dataset, additional datasets were collected for implementing the web-based prediction system. These datasets contain only the 21 optimal features selected from the primary dataset. They were gathered from students at five other colleges and include the following records,

- **Government Arts College, Panagar**: 25 student records (CSV file name: **Panagar.csv**) **–** Pass: 15, Fail: 10
- **Rajeev Gandhi Institute of Science and Technology**: 11 student records (CSV file name: **Rajeev.csv**) – Pass: 10, Fail: 1
- **Navyug Arts & Commerce College**: 31 student records (CSV file name: **Navyug.csv**) **–** Pass: 18, Fail: 13
- **St. Aloysius' College (Autonomous)**: 98 student records (CSV file name: **SAC.csv**) – Pass: 77, Fail: 21
- **Ayodhya Prasad Narmada College**: 21 student records (CSV file name: **APN.csv**) – Pass: 15, Fail: 6

**Table 3.1. Feature Detail of student_data.csv**

| Factor | Feature | F_Name | F_Description |
|---|---|---|---|
| Demographic | A1 | G | Gender |
| | A2 | A | Age |
| | A3 | M | Marital Status |
| | A4 | CHD | Number of Children |
| | A5 | SGC | Single Girl Child |
| | A6 | R | Religion |
| | A7 | NC | Number of Children's |
| | A8 | C | Category |
| Educational | A9 | AT | Admission Type |
| | A10 | CCC | Co-Curricular Activities in College |
| | A11 | ATT | Attendance |
| | A12 | FA | Father Academic Qualification |
| | A13 | MA | Mother Academic Qualification |
| | A14 | Q | Quality of Pass Out School |
| | A15 | CT | Course Type |
| | A16 | CLM | Current Learning Mode |
| | A17 | EIA | Education Importance Awareness |
| | ***A18*** | ***T*** | ***Last year grade (Target variable)*** |
| | A19 | PCS | Program Choice Satisfaction |
| | A20 | RC | Relationship with Classmates |
| | A21 | RT | Relationship with Teachers |
| | A22 | QI | Quality of Institute |
| | A23 | PI | Profession Interest |
| | A24 | CIA | Career Importance Awareness |
| | A25 | LMP | Learning Mode Preference |
| Behavioral | A26 | E | Engagement in Class |
| | A27 | CONF | Confident |
| | A28 | EMOT | Emotional |
| | A29 | MOT | Motivated |
| | A30 | BH | Bad Habits |
| | A31 | PLP | Personal Life Problems |
| | A32 | PLA | Past Life Activities |
| | A33 | HSS | Hrs. in Social Sites |
| | A34 | HVG | Hrs. in Online Video Games |
| | A35 | HC | Hrs. in Café |
| | A36 | CCH | Co-Curricular Activities in Home |
| Family | A37 | SPC | Single Parent Child |
| | A38 | FC | First Child |
| | A39 | FS | Family Support |
| | A40 | FA | Family Conflict |
| | A41 | FF | Family Freedom |
| | A42 | FMV | Family Moral Values |
| | A43 | FHT | Family Help, Trust |
| | A44 | FT | Family Type |
| | A45 | NS | Number of Siblings |
| | A46 | PMS | Parents Martial Status |

| | | | |
|---|---|---|---|
| Socioeconomic | A47 | PTJ | Part-Time Job |
| | A48 | TI | Total Income |
| | A49 | SCHO | Scholarship |
| | A50 | MO | Mother Occupation |
| | A51 | FAO | Father Occupation |
| Contextual | A52 | TRP | Transport Problem |
| | A53 | DA | Distance From College |
| | A54 | ASL | Area Student Live |
| | A55 | COV | Effect of COVID-19 Pandemic |
| | A56 | LB | Language Barrier |
| | A57 | GENI | Gender Inequality |
| | A58 | SURR | Surroundings |
| | A59 | ILLA | Involve in Illegal Activities |
| Interpersonal | A60 | FRIA | Friend Attendance |
| | A61 | RELAS | Relationship with Administrative Staff |
| | A62 | FRIB | Friend Bad Habits |
| | A63 | FRIE | Friend Engagement in Class |
| | A64 | FRIM | Friend Motivation |
| Time Management | A65 | HCOC | Hrs. in Coaching Classes |
| | A66 | HSS | Hrs. in Self-Study |
| | A67 | HLL | Hrs. in Lab, Library |
| Health | A68 | HISS | Health Issues |
| | A69 | TRA | Trauma |
| | A70 | DIS | Disability |

A combination of additional datasets was utilized to test and evaluate the prediction model's implementation via the web platform. Numerous colleges involved in data analysis demonstrate how the system provides flexible support for different student populations.

### 3.4.2 Proposed Methodology

A system that accurately predicts student academic performance is developed through implementing the methodology's defined process. All methods, including data collection, along with cleanup operations and feature selection, model training and optimization, and model evaluation, merged to deliver actionable predictions. Figure 3.2 presents the methodology flowchart that outlines the systematic process for data processing and model development. To provide a thorough explanation of the process, this section is split into two parts:

**Table 3.2. Feature Detail of stu_dataset_final.csv**

| Factor Name | Feature | F_Name |
|---|---|---|
| Demographic | A1 | Gender |
| | A2 | Age |
| | A3 | Number Of Children's |
| Educational | A4 | Co-Curricular Activities in College |
| | A5 | Attendance |
| | A6 | Father Academic Qualification |
| | A7 | Mother Academic Qualification |
| | A8 | Admission Type |
| Family | A9 | Single Parent Child |
| Contextual | A10 | Gender Inequality |
| | A11 | Transport Problem |
| Socioeconomic | A12 | Part-Time Job |
| | A13 | Father Occupation |
| | A14 | Total Income |
| Time-Management | A15 | Hours spent in Coaching classes |
| | A16 | Hours spent in Social Sites |
| | A17 | Hours spent in Online Video Games |
| | A18 | Hours spent in Cafě, canteen and campus |
| | A19 | Hours spent in Self-Study |
| Behavioral | A20 | Self-Confident |
| | A21 | Emotional Stability |
| - | *A22* | *Last year grade (Target variable)* |

**A. Optimized Classification Approach**

**a) Data Collection**

The researchers applied Cochran's (1990) formula to determine the suitable sample size for this study because this method functions for both finite and infinite population sizing. Cochran's formula calculates finite population sample size using the following expression,

$$n = \frac{n_0}{1+\frac{n_0-1}{N}} \tag{3.32}$$

where $n_0$ is calculated by using a formula,

$$n_0 = \frac{Z^2 * P(1-P)}{E^2} \tag{3.33}$$

Here, sample-size for population which is finite is n, the sample-size for population which is infinite is $n_0$, Z-score to the confidence level (95% confidence value, Z=1.959) is Z, the population proportion (supposed to be 50%) is P, the margin of error at 2% is E, and the total population size is N.

For this study, the following values were used,

- Confidence interval: 95%
- Margin of error: 2%
- Population Proportion: 50%
- Total Population: 3186

Using these parameters, the calculated sample size for an infinite population ($n_0$) was determined first. It was then adjusted for the finite population size of 3186, resulting in a final sample size of 1369.

**b) Data Preprocessing**

A crucial step in guaranteeing the dataset's quality, consistency, and analytical readiness is data preprocessing. This study employed the following sub steps in the data preprocessing phase,

- **Data Cleaning**

The dataset was cleaned to remove any inconsistencies or inaccuracies. In this study, all questions in the questionnaire were marked as required, verifying that the gathered dataset contained no missing values.

- **Data Integration**

The merging of information from various source systems, including databases, text files, and spreadsheets, into unified datasets constituted the data integration process. This research used a single Excel spreadsheet as the consistent database for record integration.

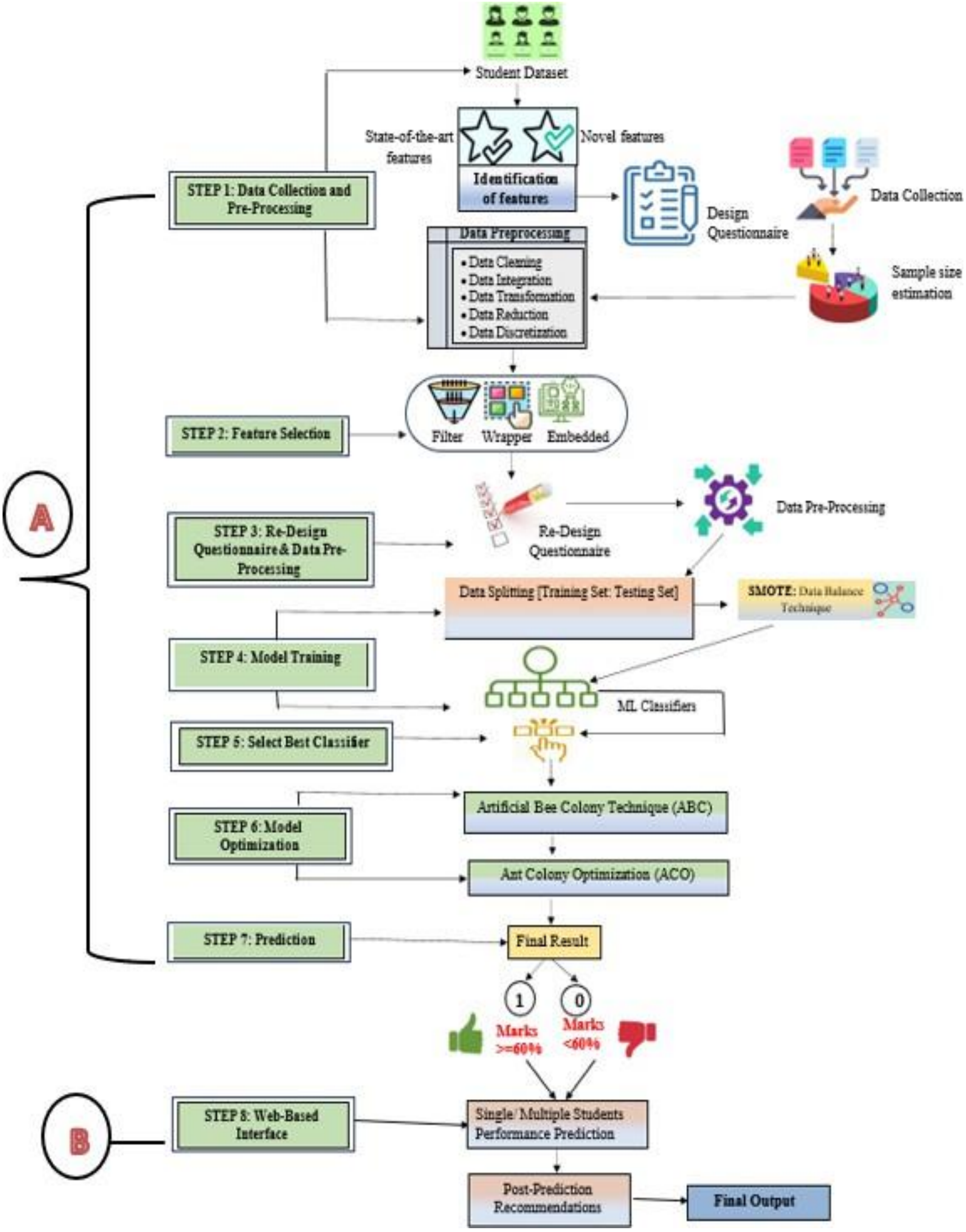


**Figure 3.2. Flowchart of the proposed methodology**

- **Data Transformation**

The appropriate formatting of data for data mining operations took place at this stage. Conversion of the Excel datasheet (.xlsx) into a CSV file format eased both data analysis steps and model integration tasks in this study.

- **Data Reduction**

To enhance the efficiency of the predictive model, irrelevant and redundant features were removed. This was achieved by performing feature selection techniques, which reduced the dataset from 70 features to 21 optimal features that significantly contributed to the prediction task.

- **Data Discretization**

The conversion of continuous numerical data to categorical data took place at this stage to enable the use of decision trees and other categorical data mining methods. The research employed label encoding as well as IFS() and IF() functions for discretization.

- **SMOTE**

To address class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied. The SMOTE was used to rectify the dataset's class imbalance. In order to balance the dataset and enhance model performance, this approach creates synthetic samples for the minority class. The SMOTE algorithm operates as follows:

1. *Step 1: Start*
2. *Step 2: Calculate the Euclidean distance between X and all other minority class records to find its k nearest neighbors.*
3. *Step 3: Randomly select one of the nearest neighbors, denoted as Xnn*
4. *Step 4: Generate a synthetic record (Xnew) using the following formula,*

$$Xnew = X + rand(0{,}1).\,Xnn - X \qquad (3.34)$$

5. *where, rand (0,1) is a random value between 0 and 1.*
6. *Step 5: End*

**Figure 3.3. SMOTE Algorithm**

Before applying SMOTE, the dataset exhibited class imbalance, with 456 records in the minority class (≤ 60%) and 913 records in the majority class (> 60%), resulting in an approximate ratio of 1:2. After applying SMOTE, both classes were balanced to 913 records each (1:1), thereby improving the model's ability to learn from both categories effectively.

**c) Feature Selection ($\chi^2$)**

The selection of important features is an essential data preparation step that focuses on identifying predictive data features. This research used chi-square tests as one of nine feature selection methods, which resulted in the most precise predictions. The Chi-Square testing procedure selected the best attributes found in the dataset. The $\chi^2$ algorithm functions according to the following procedure:

1. *Step 1: Start*
2. *Step 2: Create a contingency table for each feature that cross-tabulates the feature against the target feature. This table aids in the analysis of the connection between two features by displaying their frequency distribution.*
3. *Step 3: For each contingency table, determine the chi-square value using the formula,*

$$\chi^2 = \frac{\sum (Oi - Ei)^2}{Ei} \tag{3.35}$$

   *where the expected frequency of a feature is represented by (Ei) and the observed frequency by (Oi) in the table.*

4. *Step 4: Sort the features in descending order.*
5. *Step 5: Choose the top-k features with the highest value as a final feature set.*
6. *Step 6: End*

**Figure 3.4. Chi-Square ($\chi^2$) Algorithm**

**d) Redesign Questionnaire and Data Preprocessing**

The questionnaire underwent redesign to present the 21 main features discovered during feature selection. A selection stage used predictive success metrics for student academics to determine the analysis features. The updated questionnaire contained fewer items, focusing primarily on organizational and learning evaluation characteristics, resulting in improved data collection efficiency. Data pretreatment methods including cleaning, integration, transformation, reduction, and discretization were applied again after obtaining data from the new questionnaire. A preparation process transformed the data for upcoming evaluation and validated its consistency.

**e) Model Training and Selection**

The dataset was subjected to eight machine learning classifier tests, including DT, SVM, SVM-LK, RF, NB, XGB, LR, and KNN. Each classifier's success rate was assessed using conventional criteria in both training and testing phases. Among these eight, the

DT classifier was selected because it demonstrated the highest accuracy compared to the others, warranting further optimization and analysis. The following is the algorithm of the DT,

1. *Step 1: Start.*
2. *Step 2: Select the best feature for splitting based on the Gini Index or Information Gain.*
3. *Step 3: Create a decision node for the selected feature and divide the dataset into subsets based on its values.*
4. *Step 4: Recursively apply the splitting process to each subset until one of the following conditions is met,*
    - *All instances in a subset belong to the same class.*
    - *No further features remain for splitting.*
    - *A predefined depth limit or minimum sample size is reached.*
5. *Step 5: Assign the majority class label to leaf nodes where splitting stops.*
6. *Step 6: End.*

**Figure 3.5. Decision Tree Algorithm**

This process created a tree structure that was subsequently used to predict student academic performance. The DT's high accuracy made it reliable for deployment in the prediction system.

**f) Model Optimization**

Model optimization is essential for improving accuracy, reducing prediction time, and enhancing the generalizability of machine learning models. In this study, both ABC and ACO optimization algorithms were applied sequentially to the Decision Tree (DT) classifier, which showed the highest effectiveness during model selection. First, ABC was applied to optimize the hyperparameters, and the resulting performance was recorded. Subsequently, ACO was applied independently, and its optimized results were compared with ABC. This approach allowed evaluation of the relative effectiveness of the two techniques. ACO achieved better performance and was therefore selected for the final model.

- **ABC**

Through its foraging simulation, ABC finds optimal solutions by following honey bee natural behaviour. The research used ABC to optimize the classification procedure. Users can enhance model both performance and accuracy by identifying optimal feature subsets and parameter combinations supported by this method. The optimization process for ML model hyperparameters uses the ABC flow diagram, which appears in Figure 3.6.

Parameter initialization stands as the first step in the process, as illustrated by the figure presented below. The search area consists of parameters that encompass essential algorithm assets, including population size and iteration constraints and other such information.

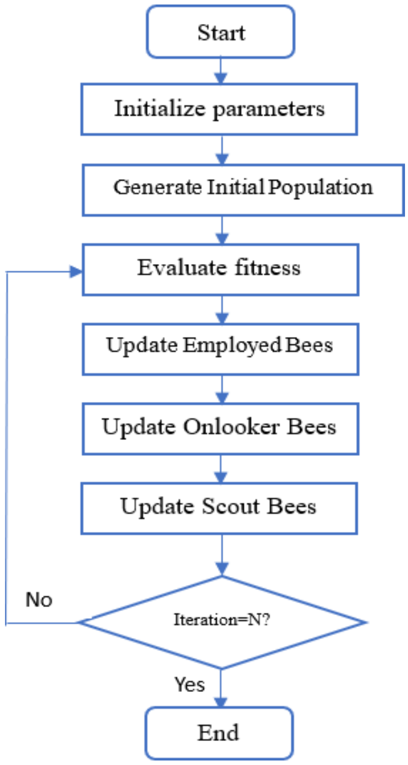


**Figure 3.6. ABC Flowchart for Hyperparameter Optimization in ML models**

When the parameters are established, the method generates a baseline population (Step 2), where each solution represents a collection of potential machine learning model hyperparameters. After that, in Step 3, each solution's performance is evaluated by training the model with the given hyperparameters and assessing the results. In Step 4, the hired bees rearrange themselves by looking for neighbouring options that are more fit-based and so perform better. The spectator bees then choose the best-performing solutions (Step 5) by updating their own based on the employed bees' solutions' fitness. After a given number of rounds, a particular solution fails to improve, therefore the scout bees are used (Step 6), where they look for completely new ideas. Following the bees' exploration of the search space and updating of their positions, the algorithm determines whether the stopping requirement has been satisfied (Step 7). The procedure concludes if the algorithm has discovered a solution that satisfies the fitness threshold or has reached the predetermined number of iterations (Step 8). If the halting condition is not satisfied, the algorithm continues the search for optimal hyperparameters by repeating the fitness assessment in Step 3.

- **ACO**

The ACO technique was based on the ant's pheromone-driven foraging method. The hyperparameters, such as maximum depth and splitting criteria, of the Decision Tree classifier were optimized in this study using ACO. As they construct potential solutions, ants follow pheromone trails that lead them to the optimal path. This iterative strategy ensures improved model performance and delivers trustworthy and accurate predictions by establishing a balance between exploration and exploitation. Figure 3.7 shows the ACO flow diagram used to optimize hyperparameters in ML models. As seen in the illustration, the first step of the ACO technique is parameter initialization. In Step 2, once the parameters have been defined, the method generates random solutions, each of which represents a potential set of hyperparameters for the machine learning model.

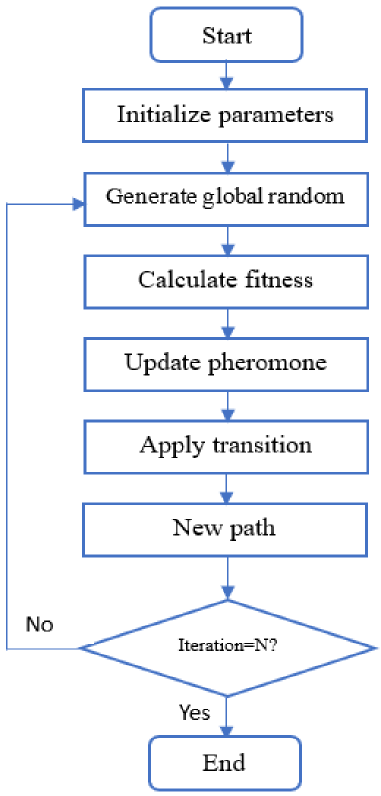


**Figure 3.7. ACO Flowchart for Hyperparameter Optimization in ML models**

In Step 3, the fitness of each solution is evaluated by analyzing the model's performance, which is often measured by accuracy or another relevant metric. Afterward, the pheromone levels are adjusted (Step 4) to reward better solutions and guide the iterations that follow towards the optimal hyperparameters. The algorithm decides whether to use the best-performing solutions or search for new ones by using a transition rule (Step 5). The algorithm finds out whether the number of iterations has reached the preset limit in Step 7, and Step 6 develops a new route. The procedure is terminated (Step 8) if the halting requirement is satisfied. Step 2 marks the algorithm's return to global random generation, which allows it to keep scanning the search space.

The study employed multiple machine learning models for predicting student performance. Among these models, only the Decision Tree (DT) was optimized using the ABC and ACO algorithms, tuning hyperparameters such as maximum depth, minimum samples split, and minimum samples per leaf. The other models, including XGB, Naive Bayes (NB), KNN, SVM, SVM-LK, Random Forest (RF), and Logistic Regression (LR), were used without optimization in this study; however, their hyperparameters can be tuned in future work. For example, number of trees, learning rate, and max depth for XGB; smoothing parameter (alpha) for NB; number of neighbors and distance metric for KNN; C, gamma, and kernel type for SVM; kernel parameters, C, and gamma for SVM-LK; number of trees, max depth, and min samples split for RF; and regularization parameter (C) and solver for LR.

#### g) Prediction and Evaluation

After the model optimization process, predictions are generated, classifying students as "Pass (1)" or "Fail (0)" based on their likelihood of achieving the specified academic threshold. The evaluation of the optimized model is performed using standard performance metrics.

### B. Web-Based Interface Development

The main motivation for creating a web-based system was to facilitate the quick and accurate estimation of a student's total score by decision-makers and academic counsellors. The system utilizes a trained ACO-DT classifier model, which had proven to be more accurate, was used by the system. An individual can enter a student's information into the system, run the student's grade prediction through the model, and then receive a PDF report outlining the findings. Figure 3.9 and Figure 3.10 shows the pseudocode and flow diagram that explain how this prediction system works.

**a) Interfaces**

The interface consisted of five web pages,

1. *Admin Login Page*: For secure admin access.
2. *Dashboard*: To select between single or multiple student predictions.
3. *Single Student Prediction Page:* For inputting data for a single student and navigating to the result page after clicking “Predict” button.
4. *Multiple Students Prediction Page*: For uploading an Excel file with multiple students' data and navigating to the result page after clicking “Predict” button.
5. *Result and Report Page:* The result and report page were a straightforward interface that displays the prediction output in the form of "Pass" or "Fail."

- **Pass:** Indicate that the student is predicted to score above or equal to 60%.
- **Fail:** Indicate that the student is predicted to score below 60%.

If the result is "Fail," the system provides tailored recommendations to help the student improve their performance.

**b) Web Server Implementation**

Django served as the web framework, while PyCharm was used as the programming environment to implement all requirements. Server-side processes became more efficient due to Django, and the project used the stable PyCharm platform for management and coding.

***Input:***

a. *Dataset D with features F and target T.*

b. *Number of ants n, maximum iterations max_iter, and pheromone evaporation rate ρ.*

***Output:***

a. *Optimized Decision Tree model M.*

1. ***Step 1: Initialize Parameters***

    a. *Set the number of ants n, max_iter, ρ, and the pheromone matrix τ for each feature split.*

    b. *Initialize the heuristic information η based on a criterion such as information gain or Gini index.*

2. ***Step 2: Start Iteration***

    *For each iteration t=1 to max_iter:*

3. ***Step 3: Ant Construction Phase***

    *For each ant i =1 to n:*

    a. *Tree Initialization: Start with an empty decision tree.*

    b. *Node Splitting: At each node, select a feature and split point probabilistically based on pheromone and heuristic information:*

$$P(f,s) = \frac{(\tau\,(f,s)^{\alpha}\cdot(\eta\,(f,s)^{\beta}}{\sum_{f' \in F, s' \in S}(\tau\,(f',s')^{\alpha})\cdot(\eta\,(f',s')^{\beta}} \qquad (3.36)$$

    *where, τ (f, s) is the pheromone level and η (f, s) is the heuristic value for feature f and split s, α, β is control the influence of pheromone and heuristic values, respectively.*

    c. *Construct Tree: Recursively split the nodes until a stopping criterion is met (e.g., maximum depth or minimum samples).*

    d. *Evaluate Tree: Calculate the fitness of the constructed tree Ti using a metric like accuracy on a validation set.*

4. ***Step 4: Pheromone Update:***

    a. *Evaporate Pheromones:*

$$\tau(f,s) \leftarrow (1-\rho)\cdot\tau(f,s) \qquad (3.37)$$

    b. *Deposit Pheromones: For each ant i, update pheromones based on the fitness fit (Ti) of the tree it constructed:*

$$\tau(f,s) \leftarrow \tau(f,s) + \Delta\tau(f,s) \qquad (3.38)$$

    *where,*

$$\Delta\tau(f,s) = \frac{fit\,(Ti)}{max\,fit} \qquad (3.39)$$

5. ***Step 5: Select Best Tree:***

    a. *Identify the tree with the highest fitness score across all iterations as the final optimized model M.*

6. ***Step 6: Terminate:***

    a. *Output the optimized Decision Tree M.*

**Figure 3.8. ACO-DT Classifier Algorithm**

```
1.  START
2.  ADMIN_LOGIN()
3.  IF login_valid THEN
            PROMPT user: "Select Prediction Type"
4.  IF user_choice = = "Single Student" THEN
           INPUT student_data
5.  IF missing_data THEN
          PROMPT: "Enter all required data"
6.  ELSE
         CALL prediction_model(student_data)
7.  ENDIF
8.  ELSE IF user_choice = = "Multiple Students" THEN
        UPLOAD file
9.  IF file_missing OR incorrect_format THEN
        PROMPT: "Upload a valid file"
10. ELSE
       CALL prediction_model(file_data)
11. ENDIF
12. ENDIF
       DISPLAY prediction_results
      GENERATE_REPORT()
13. ELSE
     PROMPT: "Invalid login credentials"
14. ENDIF
15. END
```

**Figure 3.9. Pseudocode: Web-based prediction system**

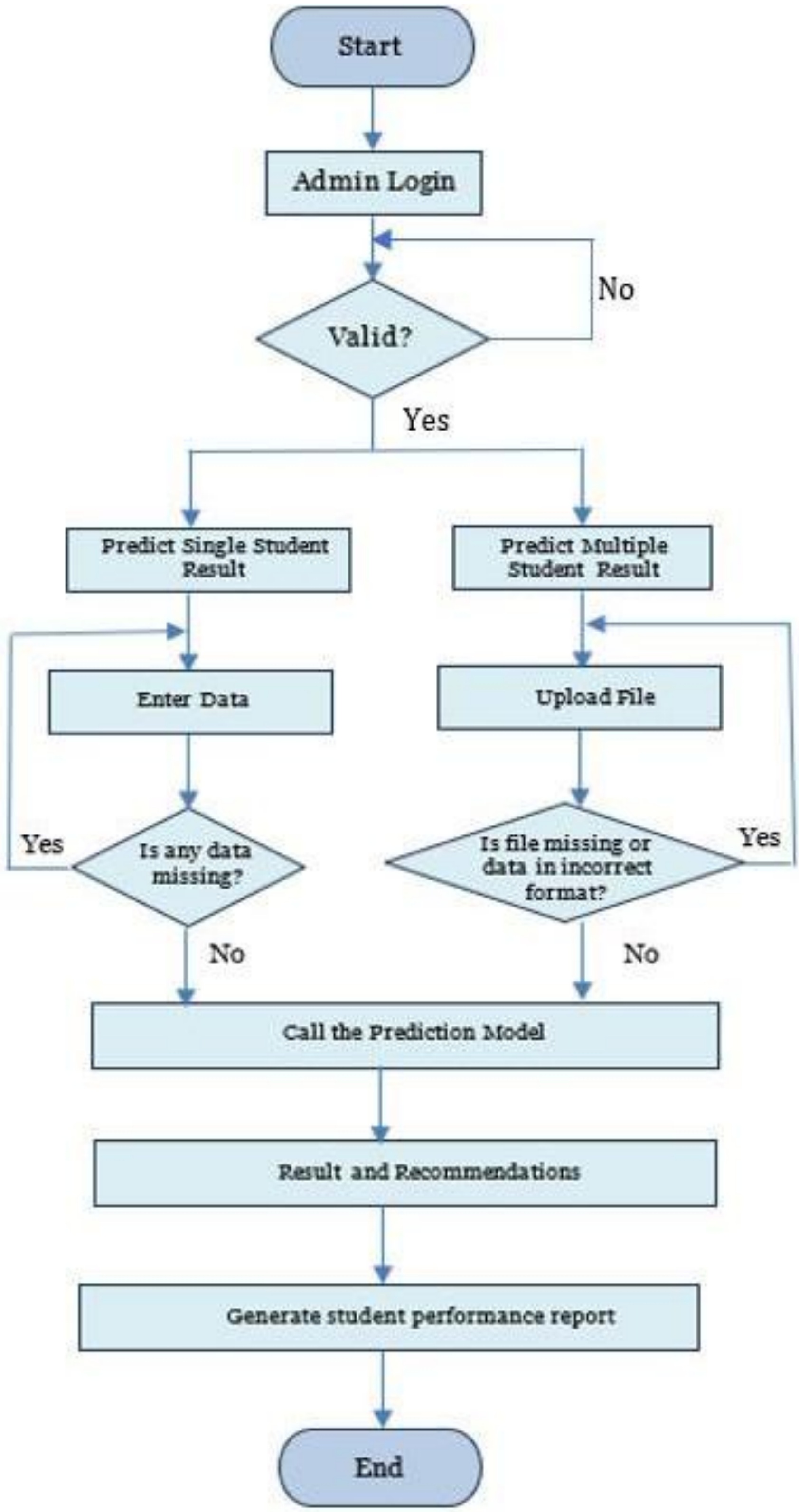


**Figure 3.10. Flowchart of the Web-based prediction system**

### 3.4.3 EXPECTED OUTCOMES

- To contribute to the existing research on forecasting student academic success by combining advanced optimization techniques with ML technique.
- The goal is to develop a robust prediction model capable of accurately classifying students as either "Pass" or "Fail," based on a rich dataset that captures a wide range of socio-economic, demographic, and academic factors.
- To optimize the performance of the Decision Tree classifier using ABC and ACO algorithms, enhancing its accuracy and reliability.
- To demonstrate the success of the proposed methodology through experimental testing using performance metrics.
- To validate the optimized prediction model on datasets from multiple institutions, ensuring its generalizability and scalability.
- To provide a web-based interface that facilitates student result forecasting and offers recommendations for weak learners.

## 3.5 SUMMARY

This chapter presents two key methodologies: the optimization-based methodology and the machine learning-based methodology, both important for developing the student academic performance prediction system. Both ABC and ACO methodologies employed in the optimization-based approach enhance the performance of the Decision Tree (DT) classifier. ABC enhances the prediction accuracy by adopting optimal hyperparameters and selecting paramount features, whereas ACO further optimizes the classifier design. Machine learning data preparation using preprocessing techniques involves recursive feature removal, Pearson correlation, and chi-square methods, along with feature extraction and selection steps. The reduction of data dimensions aids in identifying critical elements within the dataset. Preprocessing occurs before training multiple classifiers, which helps yield optimal results from the Decision Tree classifier. ABC and ACO are applied to the model to optimize its performance level. A web application, built using the combination of Python and Django, serves to host the enhanced model. Users can input student data and generate predictions, which are then presented in PDF report for analysis.

# Chapter 4
# RESULTS ANALYSIS

Machine learning techniques within student academic performance prediction help forecast whether students will achieve marks at or above a certain specified level such as 60%. The system analyzes societal, demographic, and academic factors using methods including data preprocessing, feature selection, and predictive modeling. Optimization methods such as ACO and ABC help improve the models' accuracy levels. Students who fall under the risk category can receive suitable intervention by utilizing the system, which provides forecasts through a web-based platform in real-time.

## 4.1. IMPLEMENTATION RESULTS

This segment specifies the implementation procedure for the technique while also providing assessment data from a comparison of different approaches. The development, optimization, and deployment phases of the prediction system used PyCharm, Django, and Google Colab to create both coding and web implementation. These tools expedited both the web interface verification process and the optimization of features while making models more efficient.

### 4.1.1 Feature Selection: Results and Observations

The findings preceding feature selection approaches have their own subsection, followed by a separate subsection containing the results obtained from nine feature selection processes.

#### A. Preliminary Results Without Feature Selection

Prior to the implementation of feature selection methods, this study employed the well-known logistic regression classifier to analyze the dataset. The results of this analysis are presented in Table 4.1. According to the results, the initial dataset with 70 features showed a low model fit, with 65% accuracy, 65.5% precision, 65.5% recall, a 66% f-score, and an r-squared value of 0.526. The performance metrics prior to feature selection are illustrated in Figure 4.1.

**Table 4.1. Model Performance Before FS Techniques**

| Technique | $R^2$ | Precision | Recall | F-Score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | 0.526 | 65.5% | 65.5% | 66% | 65% |


**Figure 4.1. Model Performance Metrices Prior to Feature Selection Techniques**

## B. Post-Feature Selection Results

In order to extract the most relevant features from the dataset, nine feature selection methods were used after analyzing the initial results. These methods included feature importance rankings, recursive elimination procedures, tree-based models, and methods from statistical analysis. Afterward, the logistic regression classifier was again employed to evaluate how well the selected feature subsets performed. The following subsections provide a detailed explanation of the results, including metrics for recall, accuracy, precision, F-score, and R-squared value.

### a) R-Squared Evaluation

The chi-square approach yielded an r-squared value of 0.221, representing the percentage of variance explained by the model. Compared to the earlier value of 0.526 before feature selection, this result indicates an improvement, suggesting a better fit for the model.

This result is depicted in Figure 4.2, which presents the r-squared assessment across all feature selection methods, highlighting the incremental improvements achieved through feature selection.

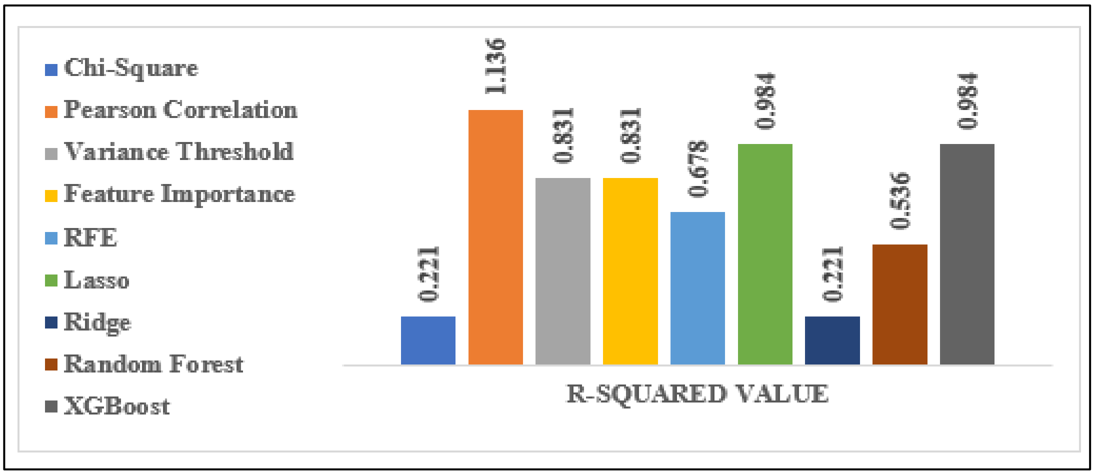


**Figure 4.2. R-Squared Evaluation**

### b) Precision Evaluation

The chi-square technique also performed well on the precision metric, which measures how well the model can cut down on erroneous positives. With the features selected using chi-square, the logistic regression classifier achieved a precision of 71.4%. Figure 4.3 provides a detailed comparison of the precision values for each method.

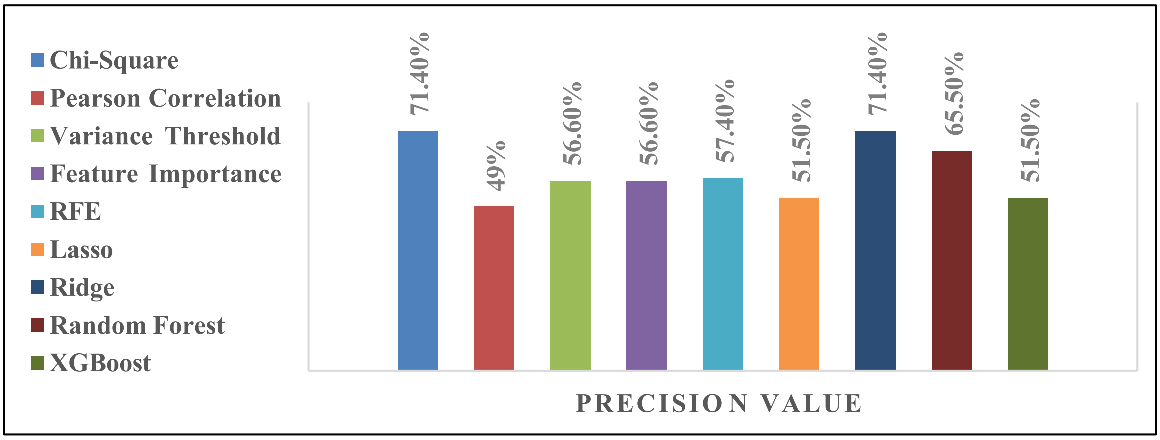


**Figure 4.3. Precision Evaluation**

### c) Recall Evaluation

The recall score, which measures the model's ability to correctly identify all relevant features, was also highest for the chi-square approach. A recall of 72.4% demonstrated the method's effectiveness in detecting true positives. Figure 4.4 shows the recall comparisons among the nine feature selection methods.

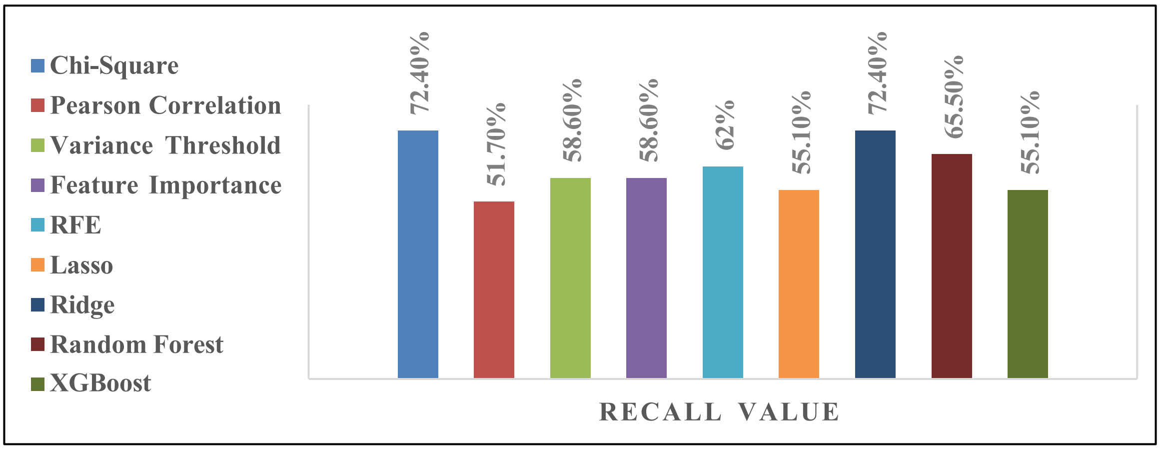


**Figure 4.4. Recall Evaluation**

### d) F-Score Evaluation

With an f-score of 70%, the chi-square approach demonstrated the best balance between recall and precision. Chi-square proves itself to be a reliable methodology for identifying optimal model features by maintaining overall performance balance. Figure 4.5 presents the f-score analysis of the nine different methods.

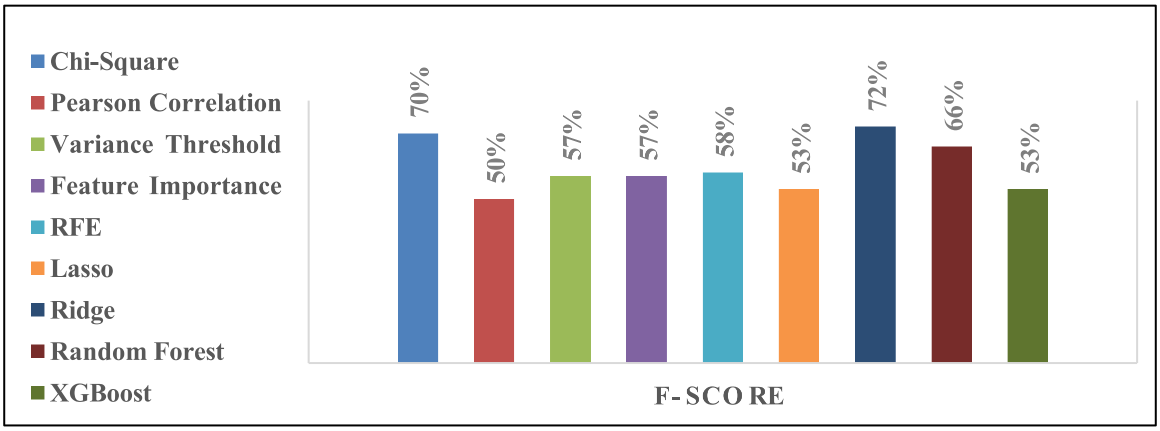


**Figure 4.5. F-Score Evaluation**

### e) Accuracy Evaluation

The logistic regression classifier, when used with features selected via the chi-square method, increased accuracy from the original dataset to 72.4%. The chi-square implementation proved to be an excellent tool for identifying the features that drive essential classification outcomes. Figure 4.6 illustrates the accuracy evaluation across all nine feature selection techniques, providing a comparative analysis of their performance.

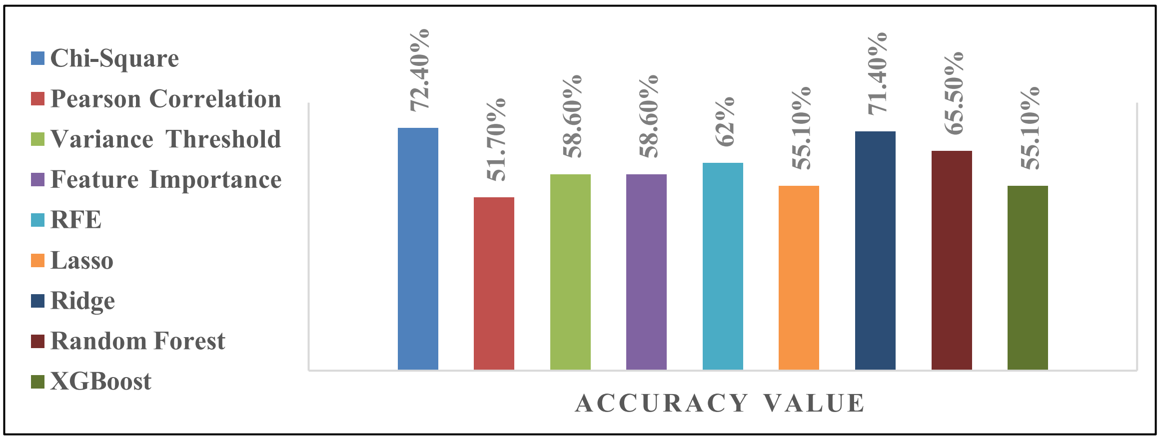


**Figure 4.6. Accuracy Evaluation**

Following the analysis of the dataset results prior to applying feature selection, nine feature selection strategies from three primary categories were utilized in this experiment. These strategies were selected from subcategories of feature selection methods, including statistical techniques, feature importance, recursive elimination, basic approaches, and tree-based methods. Among all the applied feature selection techniques, chi-square, ridge, and random forest demonstrated significant improvements in accuracy. The optimal performance metrics for each feature selection method are detailed in Table 4.2.

**Table 4.2. Comparative Analysis of the Feature Selection Techniques**

| Techniques | $R^2$ | Precision | Recall | F- Score | Accuracy |
|---|---|---|---|---|---|
| Chi-Square | 0.221 | 71.4% | 72.4% | 70% | 72.4% |
| Pearson Correlation | 1.136 | 49% | 51.7% | 50% | 51.7% |
| Variance Threshold | 0.831 | 56.6% | 58.6% | 57% | 58.6% |
| Feature Importance | 0.831 | 56.6% | 58.6% | 57% | 58.6% |
| RFE | 0.678 | 57.4% | 62% | 58% | 62% |
| Lasso | 0.984 | 51.5% | 55.1% | 53% | 55.1% |
| Ridge | 0.221 | 71.4% | 72.4% | 72% | 71.4% |
| Random Forest | 0.526 | 65.5% | 65.5% | 66% | 65.5% |
| XGBoost | 0.984 | 51.5% | 55.1% | 53% | 55.1% |

Using the logistic regression classifier, the feature subset selected by the chi-square method achieved the highest performance, with an accuracy of 72.4%, precision of 71.4%, recall of 72.4%, an f-score of 70%, and an R-squared value of 0.221. Table 4.3 summarizes the model's performance after applying the chi-square feature selection technique.

**Table 4.3. Model Performance After Applying Feature Selection Technique**

| Technique | $R^2$ | Precision | Recall | F-Score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | 0.221 | 71.4% | 72.4% | 70% | 72.4% |

Additionally, Figure 4.7 presents the model performance metrics after implementing the feature selection technique.

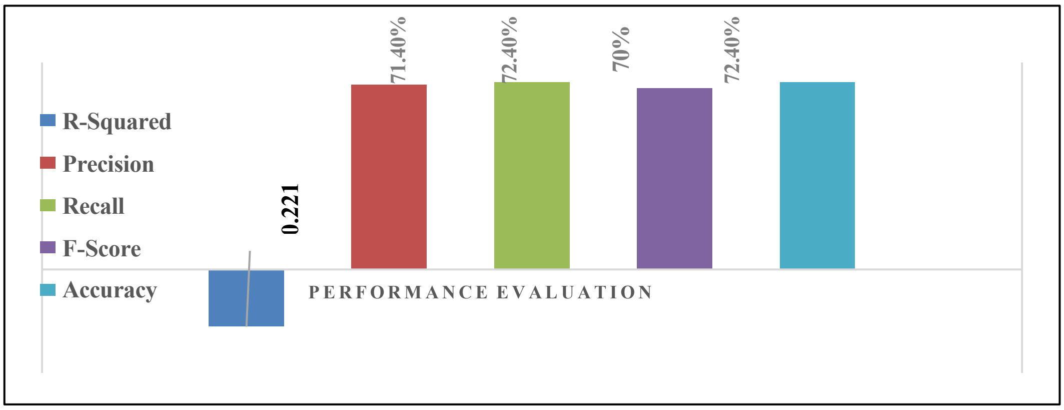


**Figure 4.7. Model Performance Metrices After Applying Feature Selection Technique**

## C. Comparative Analysis of the Model Before and After Feature Selection

Feature selection plays a crucial role in improving model performance by reducing irrelevant features and enhancing predictive accuracy. This section presents a comparative analysis of the model's performance before and after applying feature selection techniques, highlighting improvements across key evaluation metrics. Table 4.4 summarizes the performance metrics prior and post feature selection. Prior to applying feature selection, the model achieved an accuracy of 65%, with a recall and precision of 65.50% each. The R-squared value of 0.526 indicated a lower model fit. However, after implementing feature selection, particularly the chi-square method, there was a notable improvement. The model's accuracy increased to 72%, while precision and recall improved to 71.40% and 72.40%, respectively. The R-squared value also improved to 0.221, indicating a better fit and reduced error variance. The improvements in performance metrics highlight the effectiveness of feature selection in optimizing model efficiency and predictive capability.

**Table 4.4. Comparative Analysis of the Model Before and After Feature Selection**

| Performance Evaluation | R-Squared | Precision | Recall | F-Score | Accuracy |
|---|---|---|---|---|---|
| Prior to FS Techniques | 0.526 | 65.50% | 65.50% | 66% | 65% |
| After FS Techniques | 0.221 | 71.40% | 72.40% | 70% | 72% |

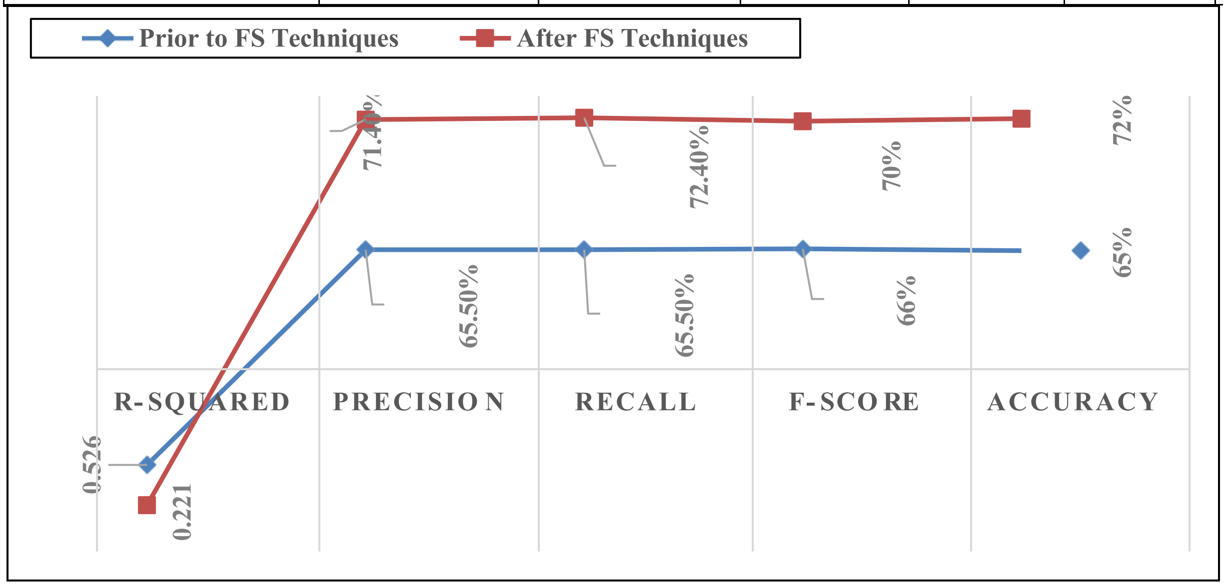


**Figure 4.8 (a). Graph representation of Model Performance Before and After Applying Feature Selection Technique**

Additionally, comparative results using the confusion matrix, ROC curve, and AUC for the model before and after applying feature selection techniques are presented in the following figures.

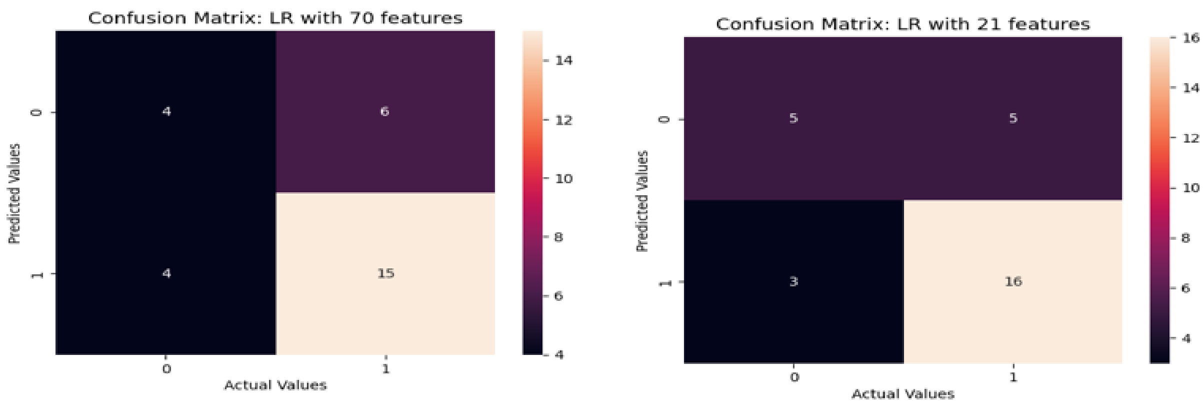

**Figure 4.8 (b). Confusion Matrix of a Model Before and After Applying Feature Selection Technique**

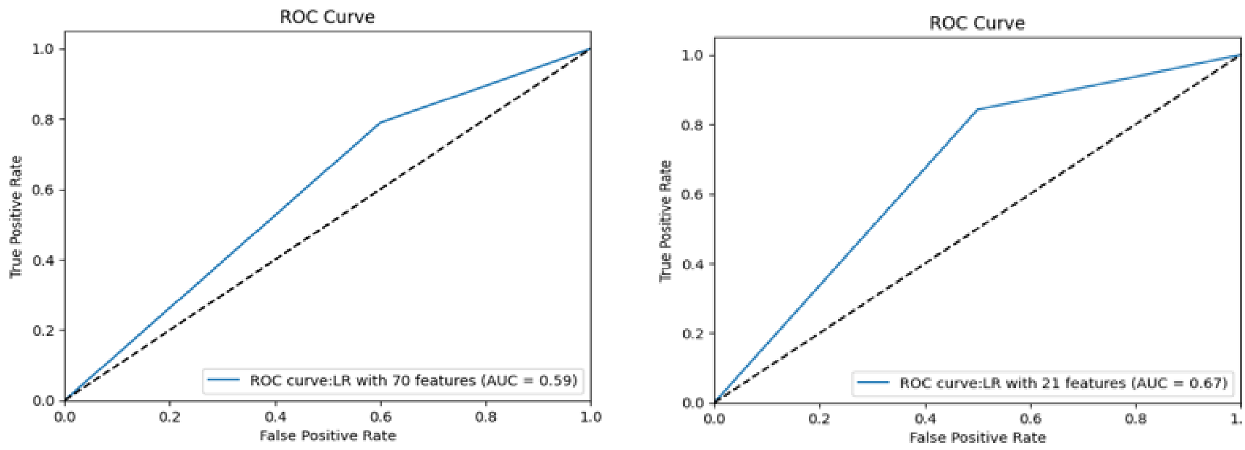


**Figure 4.8 (c). ROC Curve and AUC of a Model Before and After Applying Feature Selection Technique**

## D. Relevant Features Identified by Each Feature Selection Techniques

Each technique selected a unique subset of features based on its methodology, leading to variations in feature importance rankings. Table 4.5 presents a comparative overview of the selected features for each technique. The table indicates the presence (denoted by 1) or absence (denoted by 0) of a particular feature in the subset chosen by each method. The variations in selected features highlight the differences in how each technique evaluates feature relevance. Analyzing the results, it is evident that some features were consistently selected across multiple techniques, signifying their strong correlation with the target variable. On the other hand, certain features were selected only by specific methods, reflecting the diverse selection criteria used by different techniques.

Among the features, gender, attendance regularity, learning style, and study hours per day were found to be consistently selected by the majority of the methods, indicating their significant contribution toward predicting student performance. These features demonstrate strong relevance to academic outcomes. Conversely, 14 out of the 70 features were not selected by any of the feature selection techniques, suggesting their minimal or no predictive influence. These insignificant features include single girl child, education importance awareness, profession interest, bad consumption habits, personal life problems, past life activities, co-curricular activities in home, family type, effect of COVID-19 pandemic, language barrier, surroundings, friend motivation, health issues, and disability.

## Table 4.5. Relevant Features According to Each Feature Selection Techniques

| Features | Chi-Square | Pearson Correlation | Feature Importance | RFE | Lasso | Ridge | Random Forest | XGBoost | Variance Threshold |
|---|---|---|---|---|---|---|---|---|---|
| A1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| A2 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| A3 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| A4 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| A5 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| A6 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| A7 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| A8 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| A9 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| A10 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| A11 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| A12 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| A13 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| A14 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| A15 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| A16 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| A17 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| A18 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| A19 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| A20 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| A21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| A22 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| A23 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| A24 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| A25 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 |
| A26 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| A27 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| A28 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| A29 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| A30 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| A31 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| A32 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| A33 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| A34 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| A35 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| A36 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| A37 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| A38 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| A39 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| A40 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| A41 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| A42 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| A43 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| A44 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| A45 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| A46 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| A47 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| A48 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| A49 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| A50 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| A51 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| A52 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| A53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A54 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A55 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| A56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A57 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A58 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A59 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| A61 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A62 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A63 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A64 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A66 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A67 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A68 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A69 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

### 4.1.1 Machine Learning Techniques: Results and Observations

The evaluation of machine learning techniques was performed under different conditions, including with and without SMOTE, varying cross-validation strategies (8-fold CV and 10-fold CV), and an 80:20 train-test split. The results presented in Tables 4.6, 4.7, 4.8, and 4.9 indicate that the DT with SMOTE, an 80:20 train-test split, and 8-fold CV achieved the highest accuracy among all eight techniques.

**Table 4.6. Comparative Analysis of ML techniques with 8-fold CV (without SMOTE)**

| Techniques | R-square | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | -42.18 | 60.74% | 60.51% | 60.62% | 60.51% |
| SVM | 3.27 | 75.76% | 74.15% | 74.95% | 74.15% |
| SVM (Linear Kernal) | -30.68 | 63.87% | 63.46% | 63.66% | 63.46% |
| Random Forest | -29.30 | 64.37% | 64.33% | 64.35% | 64.33% |
| Naive Bayes | -40.20 | 61.98% | 61.84% | 61.91% | 61.84% |
| Decision Tree (Classifier) | **91.12** | **93.19%** | **95.60%** | **94.38%** | **95.60%** |
| XGBoost | **91.53** | **94.87%** | **95.60%** | **95.23%** | **95.60%** |
| KNN | **47.11** | **84.11%** | **84.12%** | **84.11%** | **84.12%** |

**Table 4.7. Comparative Analysis of ML techniques with 8-fold CV (with SMOTE)**

| Techniques | R-square | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | -45.11 | 62.40% | 63.11% | 62.75% | 63.11% |
| SVM | 5.21 | 77.82% | 75.56% | 76.67% | 75.56% |
| SVM (Linear Kernal) | -32.15 | 65.11% | 66.46% | 65.78% | 66.46% |
| Random Forest | -31.32 | 66.12% | 67.33% | 66.72% | 67.33% |
| Naive Bayes | -43.55 | 63.72% | 63.32% | 63.52% | 63.32% |
| Decision Tree (Classifier) | **93.24** | **98.93%** | **96.69%** | **97.80%** | **96.69%** |
| XGBoost | **93.13** | **96.52%** | **96.69%** | **96.60%** | **96.69%** |
| KNN | **50.14** | **86.38%** | **86.22%** | **86.30%** | **86.22%** |

The model achieved its optimal performance after implementing 8-fold cross-validation and an 80:20 train-test split, due to their ability to enhance the distribution of training and testing data and increase data subset diversity. The decision tree classifier achieved its highest prediction accuracy across all experimental results when using SMOTE in combination with the 80:20 train-test split and 8-fold CV. A comparison of different algorithm combinations is presented in the accompanying tables.

**Table 4.8. Comparative Analysis of ML techniques with 10-fold CV (without SMOTE)**

| Techniques | R-square | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | -29.77 | 54.45% | 56.34% | 55.38% | 56.34% |
| SVM | 10.34 | 72.59% | 72.43% | 72.51% | 72.43% |
| SVM (Linear Kernal) | -19.48 | 62.78% | 62.12% | 62.45% | 62.12% |
| Random Forest | -19.52 | 62.57% | 62.57% | 62.57% | 62.57% |
| Naive Bayes | -52.52 | 58.80% | 58.38% | 58.59% | 58.38% |
| Decision Tree (Classifier) | **87.42** | **91.54%** | **91.51%** | **91.52%** | **91.51%** |
| XGBoost | **87.37** | **90.42%** | **90.39%** | **90.40%** | **90.39%** |
| KNN | **33.51** | **80.78%** | **80.76%** | **80.77%** | **80.76%** |

**Table 4.9. Comparative Analysis of ML techniques with 10-fold CV (with SMOTE)**

| Techniques | R-square | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | -31.82 | 58.45% | 58.33% | 58.39% | 58.33% |
| SVM | 12.23 | 74.91% | 74.75% | 74.83% | 74.75% |
| SVM (Linear Kernal) | -21.08 | 64.81% | 64.78% | 64.79% | 64.78% |
| Random Forest | -21.15 | 64.75% | 64.72% | 64.73% | 64.72% |
| Naive Bayes | -54.85 | 60.39% | 60.34% | 60.36% | 60.34% |
| Decision Tree (Classifier) | **89.91** | **92.57%** | **92.54%** | **92.55%** | **92.54%** |
| XGBoost | **89.89** | **92.72%** | **90.39%** | **92.70%** | **90.39%** |
| KNN | **35.55** | **82.82%** | **82.76%** | **82.80%** | **82.76%** |

As the model attained optimal performance following the application of 8-fold cross-validation, SMOTE, and an 80:20 train-test split, the confusion matrix, ROC curve, and AUC of the best-performing model are illustrated in the following figures.

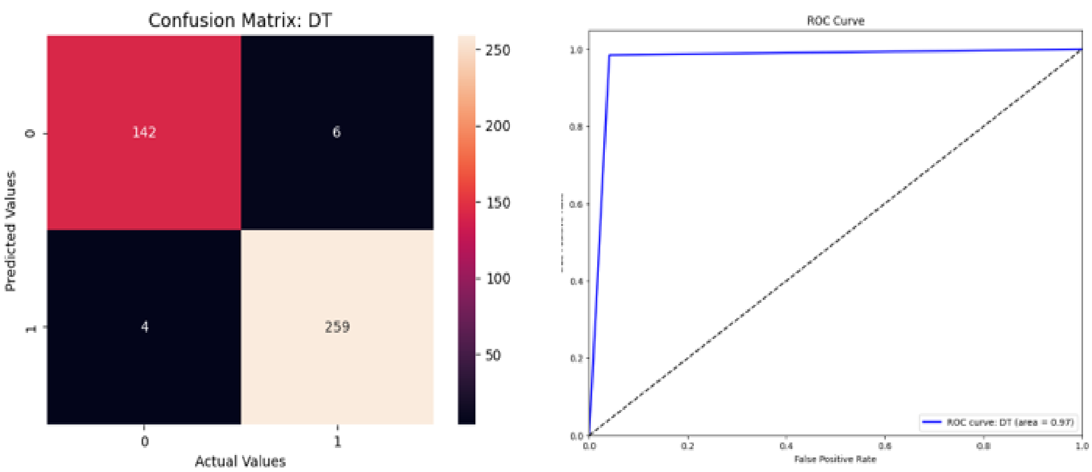


**Figure 4.9. An Illustration of the DT Classifier's Performance Result**

### 4.1.2 Optimization Techniques: Results and Observations

As the results illustrate that the Decision Tree classifier outperformed all other classifiers, this section focuses on optimizing DT using optimization techniques.

#### A. ABC-DT

ABC-DT leverages the Artificial Bee Colony (ABC) algorithm to optimize decision tree hyperparameters. This optimization process is inspired by the honeybee's foraging behavior, allowing the model to exploit, expand and explore the search space efficiently. The impact of SMOTE, cross-validation methods, and train-test splits on ABC-DT performance is detailed in Table 4.10 and Figure 4.9 below.

**Table 4.10. Performance Results of the ABC-DT Classifier**

| Validation Technique | SMOTE | F1-Score | Accuracy | Precision | Recall | $R^2$ |
|---|---|---|---|---|---|---|
| **8-fold CV** | Without SMOTE | 97.12% | 97.24% | 97.10% | 97.12% | 84.30 |
| **8-fold CV** | **With SMOTE** | **97.68%** | **97.75%** | **97.72%** | **97.68%** | **84.50** |
| **10-fold CV** | Without SMOTE | 97.05% | 97.12% | 97.08% | 97.05% | 84.10 |
| **10-fold CV** | With SMOTE | 97.50% | 97.62% | 97.58% | 97.50% | 84.35 |

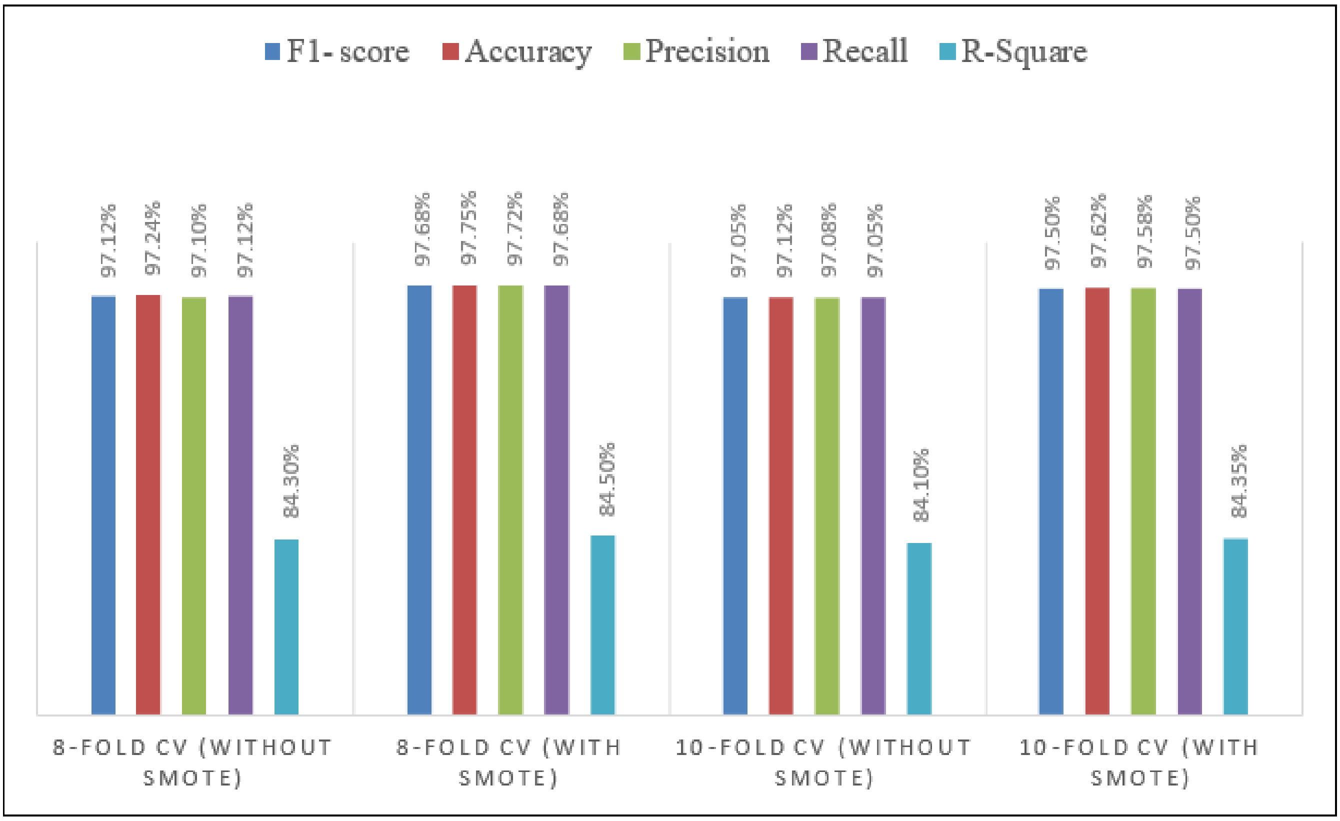


**Figure 4.9. An Illustration of the ABC-DT Classifier's Performance Result**

After applying the ABC optimization technique, the following hyperparameters (Table 4.11) were identified for decision tree training,

**Table 4.11. Optimized Hyperparameters of the DT Classifier from ABC**

| SMOTE Status | Max_depth | Min_samples_leaf | Min_samples_split |
|---|---|---|---|
| Without SMOTE | 19 | 1 | 3 |
| With SMOTE | 17 | 2 | 6 |

Key Observations for ABC-DT,

- Using SMOTE improved overall accuracy, precision, and recall across both cross-validation techniques.
- The 8-fold cross-validation setup produced slightly better accuracy than 10-fold CV, showing its effectiveness in training stability.
- Without SMOTE, the model favoured deeper trees with lower sample splits, whereas SMOTE led to slightly shallower trees with better generalization.

## B. ACO-DT

ACO-DT applies the Ant Colony Optimization (ACO) technique to optimize decision tree hyperparameters. ACO simulates the behavior of ants searching for the shortest path between their colony and a food source, where pheromone trails guide the selection of optimal solutions. This method enables the model to efficiently determine the best hyperparameter configurations for improved performance (Figure 4.10). Table 4.12 and Figure 4.11 depicts the ACO-DT classifier performance.

```
Cross Validation Scores: [0.96956522 0.98695652 0.98695652 0.98695652 0.97826087 0.96956522
 0.99563319 0.97816594]
Average CV Score: 0.9815074995253464
R SQUARE: 0.847754
PRECISION: 0.9624747622643837
RECALL: 0.9619565217391305
CLASSIFICATION REPORT:
              precision    recall  f1-score   support

           0       0.95      0.98      0.96       180
           1       0.98      0.95      0.96       188

    accuracy                           0.96       368
   macro avg       0.96      0.96      0.96       368
weighted avg       0.96      0.96      0.96       368
```

**Figure 4.10. Performance Result of the ACO-DT Classifier**

**Table 4.12. Performance Results of the ACO-DT Classifier**

| Validation Technique | SMOTE | F1-score | Accuracy | Precision | Recall | $R^2$ |
|---|---|---|---|---|---|---|
| **8-fold CV** | Without SMOTE | 97.78% | 97.84% | 97.81% | 97.78% | 84.98 |
| **8-fold CV** | **With SMOTE** | **97.16%** | **98.15%** | **98.18%** | **98.15%** | **84.75** |
| **10-fold CV** | Without SMOTE | 97.56% | 97.68% | 97.62% | 97.56% | 84.50 |
| **10-fold CV** | With SMOTE | 97.98% | 97.92% | 98% | 97.98% | 84.62 |

After applying the ACO optimization technique, the following optimal hyperparameters were identified (Table 4.13),

**Table 4.13. Optimized Hyperparameters of the DT Classifier from ACO**

| SMOTE Status | Max_depth | Min_samples_leaf | Min_samples_split |
|---|---|---|---|
| **Without SMOTE** | **11** | **4** | **2** |
| **With SMOTE** | **18** | **1** | **6** |

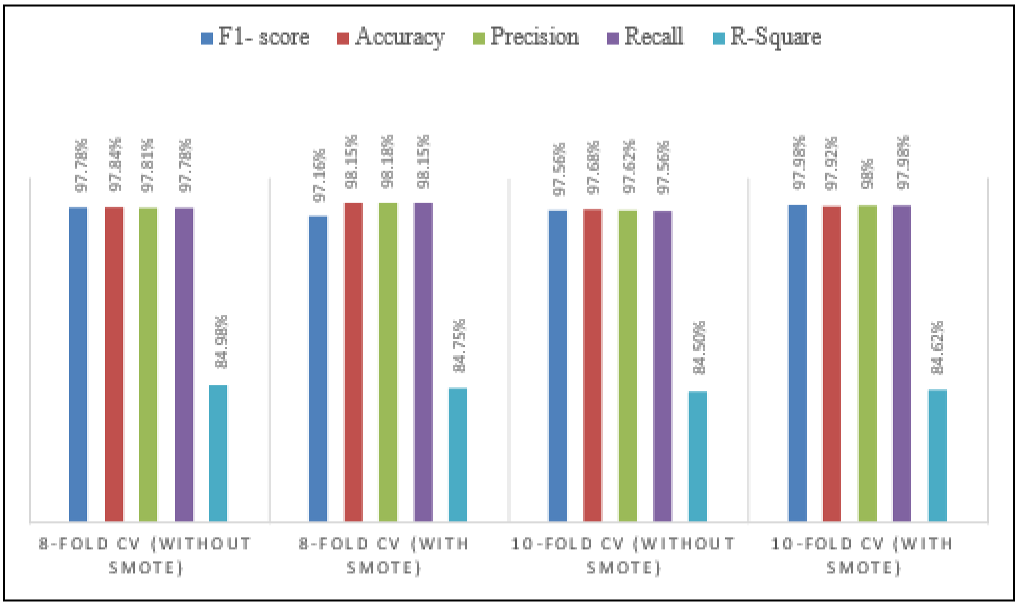


**Figure 4.11. An Illustration of the ACO-DT Classifier's Performance Result**

Key Observations for ACO-DT,

- The highest accuracy (98.15%) was achieved using SMOTE with 8-fold CV.
- Applying SMOTE resulted in deeper trees (higher max depth), which suggests better model generalization with more complex decision boundaries.

- Compared to ABC-DT, ACO-DT produced slightly better overall accuracy, indicating its superior hyperparameter tuning capabilities.
- Using 8-fold CV instead of 10-fold CV yielded better results, confirming its effectiveness in optimization tasks.

In conclusion, as shown in Figure 4.12, ACO-DT consistently outperformed ABC-DT in terms of all performance metrices, demonstrating its effectiveness in optimizing decision tree performance using 8-fold cross-validation with SMOTE. The application of SMOTE further enhanced model performance across both techniques, particularly by improving recall and accuracy through class imbalance handling. The optimized hyperparameters varied, with ACO-DT selecting deeper trees for better generalization, while ABC-DT resulted in shallower structures with different split criteria. The best-performing configuration was ACO-DT with SMOTE, 8-fold CV, and an 80:20 train-test split, achieving an accuracy of 98.15%. These findings emphasize the critical role of feature selection, SMOTE, and hyperparameter optimization in improving decision tree classification performance.

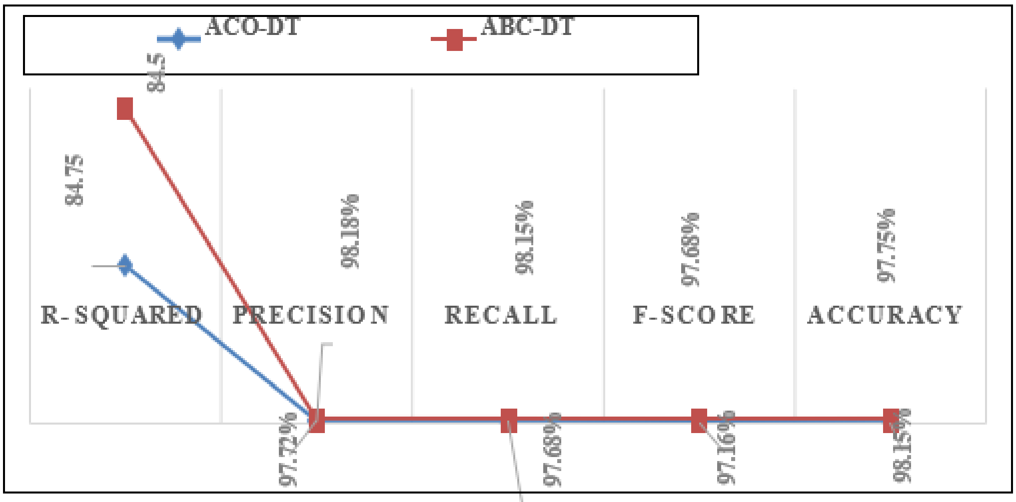


**Figure 4.12 (a). Graph Representation of ACO-DT and ABC-DT Model Performance**

Additionally, comparative results using the confusion matrix, ROC curve, and AUC for the ACO-DT and ABC-DT models are presented in the following figures.

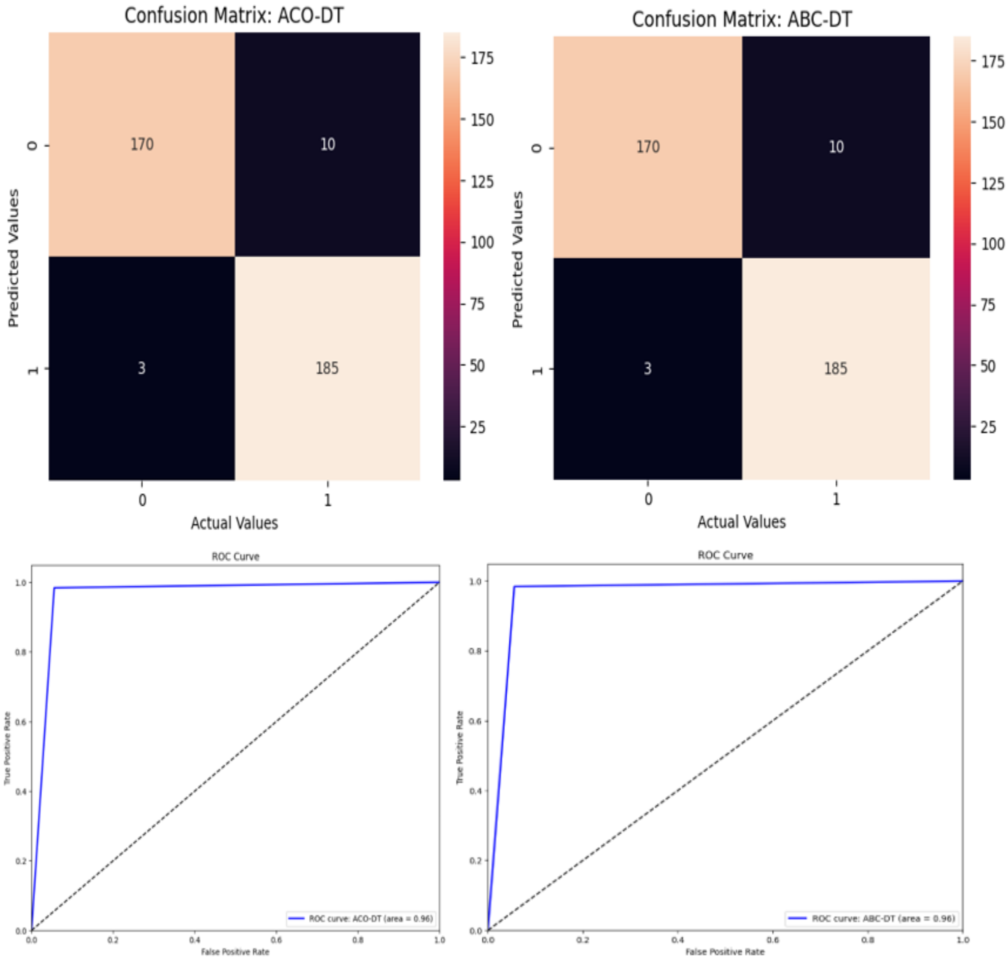


**Figure 4.12 (b). Confusion Matrix, ROC Curve and AUC of the ACO-DT and ABC-DT Model**

## C. Comparative Analysis of the Proposed Model with state-of-the-art models

The proposed ACO-DT classifier model is also compared with state-of-the-art existing models. Below is the comparison results table of the model, based on various performance metrics.

**Table 4.14 (a). Comparative Analysis of the ACO-DT model with state-of-the-art models**

| Ref. | Techniques | Dataset and Data Source | Class Balancing | Feature Selection | Features Used | Ensemble Method | Model Optimization |
|---|---|---|---|---|---|---|---|
| [4] | DT (Classifier) | 1044 data from UCI Repository | Yes | Yes | 12 features (Demographic, academic and family factors) | Yes | No |
| [16] | DT (Classifier) | 80,000 data from Islamabad, Pakistan | No | Yes | 29 features (Demographic, academic and socioeconomic factors) | Yes | No |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [53] | LR (Classifier) | 137 students from India and Ethiopia | No | No | 8 features (Demographic, family and socioeconomic factors) | No | No |
| [54] | HCSO-PSO-RNN (Classifier) | 1001 data from Kaggle Repository | No | Yes | 7 features (Demographic, academic and socioeconomic factors) | No | Yes |
| [69] | VR-ODRNN (Classifier) | 145 data from Kaggle Repository | No | Yes | 8 features (Demographic, academic and Time-management factors) | No | Yes |
| [71] | SVM (Regressor) | 20,638 data from Somaliland | No | No | 4 features (Demographic, socioeconomic and geographic factors) | No | No |
| [134] | Multiple Linear Regression (Regressor) | 1157 data from universities | No | Yes | 4 features (Demographic and academic factors) | No | No |
| [153] | Multiple Linear Regression (Regressor) | 353 data from Shenzhen (China) | No | Yes | 15 features (Demographic, academic and Behavioral factors) | No | No |
| [184] | RF (Regressor) | 168 data from IAU, Saudi Arabia | No | No | 10 features (Demographic and academic factors) | No | No |
| [192] | Ensemble - Voting (Classifier) | 480 data from Kalboard 360 | No | Yes | 16 features (Demographic, academic and socioeconomic factors) | Yes | Yes |
| **Our Proposed Model** | **ACO-DT** | **1369 data from primary source** | **Yes** | **Yes** | **21 features from 9 factors** | **No** | **Yes** |

**Table 4.14 (b). Comparative Analysis of the ACO-DT model with state-of-the-art models**

| Ref. | R-square | Precision | Recall | F1-score | Accuracy | MAPE (Mean Absolute Percentage Error) | MCC (Matthews Correlation Coefficient) |
|---|---|---|---|---|---|---|---|
| [4] | No | No | No | No | 95.78% | N/A | No |
| [16] | 8.23 | No | No | No | 96.64% | N/A | No |
| [53] | No | No | No | No | 72% | N/A | No |
| [54] | No | 96.2% | 96.4% | No | 97.9% | N/A | No |
| [69] | No | 87.88% | 78.56% | 80.61% | 96.29% | N/A | 79.93% |
| [71] | No | No | No | No | No | 9.29% | No |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [134] | 36.78 | No | No | No | No | 3.38% | No |
| [153] | No | No | No | No | No | 5.53% | No |
| [184] | No | No | No | No | No | 6.34% | No |
| [192] | No | 87% | 87% | 87% | 86% | N/A | No |
| **Our Proposed Model** | **84.75** | **98.15%** | **98.18%** | **97.16%** | **98.15%** | **N/A** | **95.12%** |

**MAPE is applicable only for regression models. For classification models, this metric is marked as 'N/A'.*

### 4.1.3 Feature Importance and Relevancy Analysis: Results and Observations

To address the first objective, this study aimed to determine the most influential features contributing to students' academic achievement. The Chi-Square Test was used to compute feature importance scores, as it is an effective statistical method for measuring the dependency between categorical features and the target variable. Feature analysis was conducted using feature relevance assessment, correlation analysis, and multiple linear regression to identify highly correlated features.

#### A. Feature Relevance Analysis Using the Chi-Square

Features with higher relevancy scores were considered more impactful in determining whether students would score above or below 60%. The analysis revealed that gender, age, and part-time job were the top three most influential features, followed by other contributing factors. Figure 4.13 presents the relevancy scores of all features, highlighting the most significant contributors.

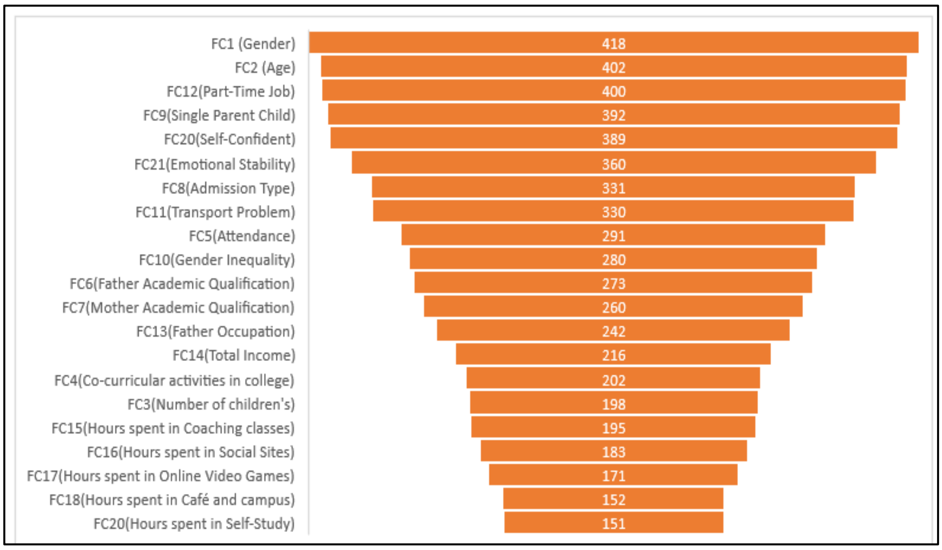


**Figure 4.13. Features Relevancy Score**

These insights are essential for designing targeted interventions and providing personalized recommendations, ultimately improving students' academic outcomes.

### B. Correlation Analysis Between Features Using the Kendall-Tau Method

The strength of relationships between feature pairs was assessed by calculating correlation coefficients using the Kendall-Tau method. Additionally, a feature correlation heatmap is provided for better visualization (Figure 4.14). This analysis helped identify features with no relationship, negative correlations (indicating inverse relationships), and positive correlations (where values increase or decrease together). The results of this analysis are presented in Table 4.15.

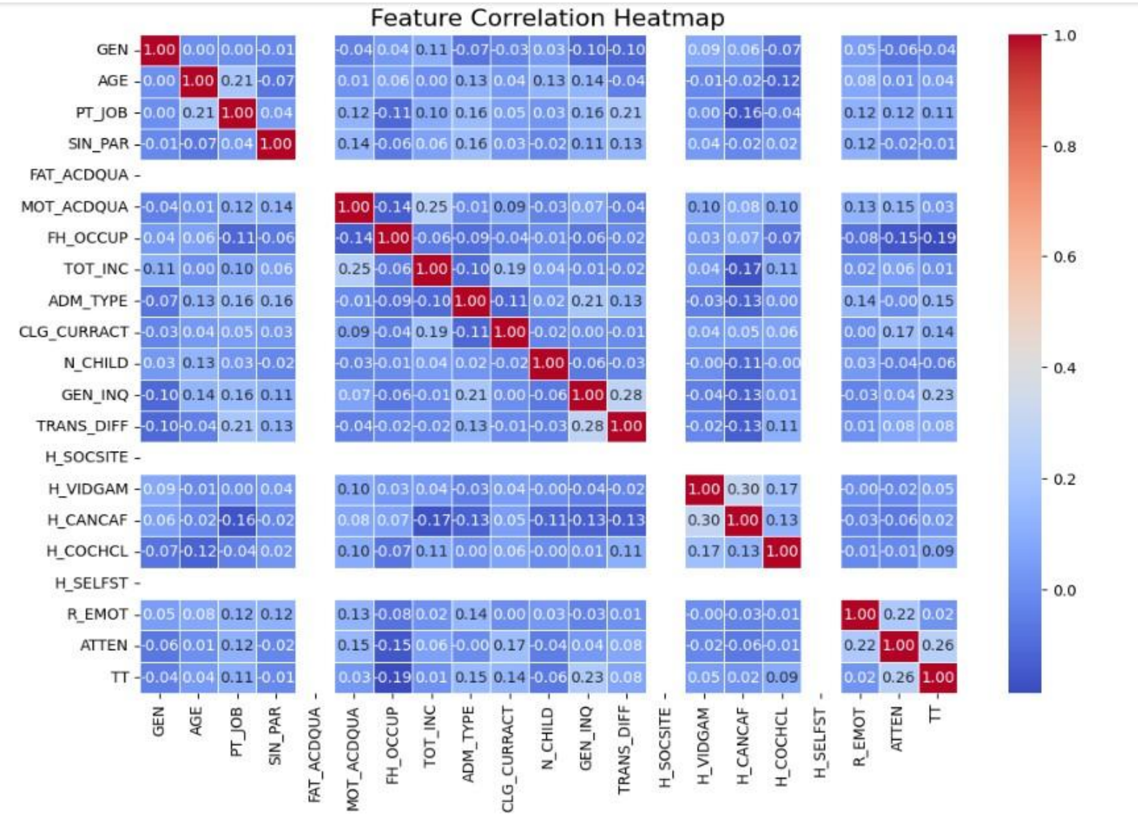


**Figure 4.14. Feature Correlation Heatmap**

#### a) Positive Correlation with the Target Feature

Features that have a positive correlation (> 0) with T (the target feature) indicate that as their values increase, the likelihood of scoring 60% or higher also increases. Table 4.16 presents these features.

#### b) Negative Correlation with the Target Variable

Features with a negative correlation (< 0) with T indicate that as these feature values increase, the likelihood of scoring below 60% also increases. Table 4.17 presents these features.

**Table 4.15. Features Correlation Analysis**

| Features | G | A | PTJ | SPC | FA | MA | FAO | TI |
|---|---|---|---|---|---|---|---|---|
| G | 1 | -0.005 | 0.001 | -0.006 | 0.081 | -0.028 | 0.039 | 0.11 |
| A | -0.005 | 1 | 0.197 | -0.073 | 0.037 | 0.016 | 0.056 | 0.009 |
| PTJ | 0.001 | 0.197 | 1 | 0.037 | 0.08 | 0.104 | -0.1 | 0.068 |
| SPC | -0.006 | -0.073 | 0.037 | 1 | 0.117 | 0.125 | -0.056 | 0.06 |
| FA | 0.081 | 0.037 | 0.08 | 0.117 | 1 | 0.455 | -0.189 | 0.221 |
| MA | -0.028 | 0.016 | 0.104 | 0.125 | 0.455 | 1 | -0.133 | 0.217 |
| FAO | 0.039 | 0.056 | -0.1 | -0.056 | -0.189 | -0.133 | 1 | -0.042 |
| TI | 0.11 | 0.009 | 0.068 | 0.06 | 0.221 | 0.217 | -0.042 | 1 |
| AT | -0.079 | 0.113 | 0.16 | 0.161 | -0.043 | -0.016 | -0.079 | -0.083 |
| CCC | -0.021 | 0.034 | 0.048 | 0.031 | 0.026 | 0.097 | -0.045 | 0.157 |
| NC | 0.049 | 0.104 | 0.026 | -0.027 | 0.017 | -0.011 | -0.012 | 0.048 |
| GENI | -0.111 | 0.125 | 0.156 | 0.11 | 0.038 | 0.063 | -0.059 | -0.019 |
| TRP | -0.103 | -0.043 | 0.205 | 0.134 | -0.055 | -0.05 | -0.008 | -0.043 |
| HSS | -0.034 | 0.03 | 0.022 | 0.009 | 0.096 | 0.097 | -0.042 | 0 |
| HVG | 0.092 | -0.021 | 0.02 | 0.062 | 0.147 | 0.108 | 0.018 | 0.059 |
| HC | 0.028 | -0.02 | -0.128 | -0.023 | 0.109 | 0.066 | 0.042 | -0.061 |
| HCOC | -0.048 | -0.12 | -0.037 | 0.053 | 0.047 | 0.082 | -0.065 | 0.107 |
| HSS | -0.065 | 0.037 | -0.044 | -0.105 | 0.016 | 0.031 | 0.084 | -0.014 |
| CONF | -0.065 | 0.021 | -0.022 | -0.006 | 0.055 | 0.117 | -0.055 | -0.029 |
| EMOT | 0.05 | 0.077 | 0.1 | 0.118 | 0.099 | 0.107 | -0.067 | 0.005 |
| ATT | -0.062 | 0.003 | 0.126 | -0.037 | 0.071 | 0.134 | -0.146 | 0.033 |
| T | -0.042 | 0.032 | 0.112 | -0.009 | 0.037 | 0.044 | -0.172 | -0.013 |

**Table 4.15. Features Correlation Analysis (continue)**

| Features | AT | CCC | NC | GENI | TRP | HSS | HVG |
|---|---|---|---|---|---|---|---|
| G | -0.079 | -0.021 | 0.049 | -0.111 | -0.103 | -0.034 | 0.092 |
| A | 0.113 | 0.034 | 0.104 | 0.125 | -0.043 | 0.03 | -0.021 |
| PTJ | 0.16 | 0.048 | 0.026 | 0.156 | 0.205 | 0.022 | 0.02 |
| SPC | 0.161 | 0.031 | -0.027 | 0.11 | 0.134 | 0.009 | 0.062 |
| FA | -0.043 | 0.026 | 0.017 | 0.038 | -0.055 | 0.096 | 0.147 |
| MA | -0.016 | 0.097 | -0.011 | 0.063 | -0.05 | 0.097 | 0.108 |
| FAO | -0.079 | -0.045 | -0.012 | -0.059 | -0.008 | -0.042 | 0.018 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **TI** | -0.083 | 0.157 | 0.048 | -0.019 | -0.043 | 0 | 0.059 |
| **AT** | 1 | -0.103 | 0.03 | 0.206 | 0.132 | -0.055 | -0.026 |
| **CCC** | -0.103 | 1 | -0.016 | 0.002 | -0.007 | 0.05 | 0.063 |
| **NC** | 0.03 | -0.016 | 1 | -0.069 | -0.028 | 0.021 | -0.003 |
| **GENI** | 0.206 | 0.002 | -0.069 | 1 | 0.284 | -0.06 | -0.039 |
| **TRP** | 0.132 | -0.007 | -0.028 | 0.284 | 1 | -0.11 | -0.014 |
| **HSS** | -0.055 | 0.05 | 0.021 | -0.06 | -0.11 | 1 | 0.068 |
| **HVG** | -0.026 | 0.063 | -0.003 | -0.039 | -0.014 | 0.068 | 1 |
| **HC** | -0.106 | 0.053 | -0.12 | -0.099 | -0.088 | 0.107 | 0.284 |
| **HCOC** | 0.012 | 0.071 | -0.003 | 0.013 | 0.121 | -0.106 | 0.227 |
| **HSS** | -0.059 | -0.063 | 0.053 | -0.004 | -0.035 | 0.074 | -0.25 |
| **CONF** | -0.029 | 0.076 | -0.023 | 0.006 | -0.018 | 0.036 | -0.067 |
| **EMOT** | 0.118 | 0.001 | 0.052 | -0.023 | 0.01 | 0.062 | -0.023 |
| **ATT** | 0.007 | 0.16 | -0.046 | 0.041 | 0.09 | -0.02 | -0.04 |
| **T** | 0.154 | 0.125 | -0.075 | 0.236 | 0.091 | -0.06 | 0.069 |

## Table 4.15. Features Correlation Analysis (continue)

| Features | HC | HCOC | HSS | CONF | EMOT | ATT | T |
|---|---|---|---|---|---|---|---|
| **G** | 0.028 | -0.048 | -0.065 | -0.065 | 0.05 | -0.062 | -0.042 |
| **A** | -0.02 | -0.12 | 0.037 | 0.021 | 0.077 | 0.003 | 0.032 |
| **PTJ** | -0.128 | -0.037 | -0.044 | -0.022 | 0.1 | 0.126 | 0.112 |
| **SPC** | -0.023 | 0.053 | -0.105 | -0.006 | 0.118 | -0.037 | -0.009 |
| **FA** | 0.109 | 0.047 | 0.016 | 0.055 | 0.099 | 0.071 | 0.037 |
| **MA** | 0.066 | 0.082 | 0.031 | 0.117 | 0.107 | 0.134 | 0.044 |
| **FAO** | 0.042 | -0.065 | 0.084 | -0.055 | -0.067 | -0.146 | -0.172 |
| **TI** | -0.061 | 0.107 | -0.014 | -0.029 | 0.005 | 0.033 | -0.013 |
| **AT** | -0.106 | 0.012 | -0.059 | -0.029 | 0.117 | 0.007 | 0.154 |
| **CCC** | 0.053 | 0.071 | -0.063 | 0.076 | 0.001 | 0.16 | 0.125 |
| **NC** | -0.12 | -0.003 | 0.053 | -0.023 | 0.052 | -0.046 | -0.075 |
| **GENI** | -0.099 | 0.013 | -0.004 | 0.006 | -0.023 | 0.041 | 0.236 |
| **TRP** | -0.088 | 0.121 | -0.035 | -0.018 | 0.01 | 0.09 | 0.091 |
| **HSSC** | 0.107 | -0.106 | 0.074 | 0.036 | 0.062 | -0.02 | -0.06 |
| **HVG** | 0.284 | 0.227 | -0.25 | -0.067 | -0.023 | -0.04 | 0.069 |
| **HC** | 1 | 0.186 | -0.037 | 0.04 | -0.069 | -0.026 | 0.083 |
| **HCOC** | 0.186 | 1 | -0.061 | 0.191 | -0.014 | 0.01 | 0.09 |
| **HSS** | -0.037 | -0.061 | 1 | 0.242 | 0.082 | 0.047 | 0.087 |
| **CONF** | 0.04 | 0.191 | 0.242 | 1 | 0.137 | 0.115 | 0.162 |
| **EMOT** | -0.069 | -0.014 | 0.082 | 0.137 | 1 | 0.216 | 0.042 |
| **ATT** | -0.026 | 0.01 | 0.047 | 0.115 | 0.216 | 1 | 0.28 |

| T | 0.083 | 0.09 | 0.087 | 0.162 | 0.042 | 0.28 | 1 |
|---|---|---|---|---|---|---|---|

**Table 4.16. Positive Correlated Features**

| Features | Correlation with T | Interpretation |
|---|---|---|
| **ATT** | 0.28 | Higher attendance is strongly associated with better academic performance, as regular class participation enhances learning and engagement. |
| **GENI** | 0.236 | Students experiencing gender-related challenges may exert extra effort to succeed. |
| **AT** | 0.154 | Certain admission types are associated with higher academic performance, possibly due to selection criteria or student preparedness. |
| **CONF** | 0.162 | Higher confidence is linked to better academic outcomes due to increased engagement and motivation. |
| **TI** | 0.107 | Financial stability offers better access to academic resources, leading to improved performance. |
| **TRP** | 0.091 | Overcoming transport-related challenges may indicate discipline and determination. |
| **HCOC** | 0.09 | Coaching classes provide additional academic support, improving understanding and performance. |
| **HSS** | 0.087 | More self-study time is associated with better academic performance. |
| **HC** | 0.083 | Weak positive correlation suggests informal learning environments like cafés/canteens may enhance academic performance through peer discussions. |
| **MA** | 0.044 | Students whose mothers have higher education levels may receive better academic support at home. |
| **EMOT** | 0.042 | Emotionally stable students may manage stress better, leading to slightly better performance. |
| **A** | 0.032 | Older students may have better time management and discipline. |

**Table 4.17. Negative Correlated Features**

| Features | Correlation with T | Interpretation |
|---|---|---|
| **SPC** | -0.009 | Minimal negative effect, but students from single-parent households may face additional challenges. |
| **HSSC** | -0.06 | Increased social media usage is linked to reduced academic engagement and lower performance. |

| | | |
|---|---|---|
| **CCC** | -0.063 | Excessive participation in extracurricular activities may reduce study time, impacting academic success. |
| **HVG** | -0.067 | More gaming time may reduce study hours, negatively impacting academic results. |
| **FAO** | -0.172 | High-status jobs may reduce parental academic involvement, affecting student performance. |

### c) No Significant Correlation with the Target Variable

Features with near-zero correlation to T suggest that they have minimal direct impact on academic performance. Table 4.18 presents these features. These findings suggest that gender, father's education, part-time jobs, and family size do not have a very much influence on student's success rate.

**Table 4.18. Features with No Correlation with T**

| Features | Correlation with T | Interpretation |
|---|---|---|
| **G** | -0.042 | Weak negative correlation suggests that gender does not significantly impact academic success. |
| **NC** | -0.075 | A weak negative correlation suggests that family size has little impact on academic success. |
| **FA** | 0.037 | A near-zero positive correlation suggests that the father's education level does not significantly impact student performance. |
| **PTJ** | 0.112 | A weak positive correlation suggests that working part-time may improve discipline but does not directly impact academic performance. |

### d) Multiple Linear Regression for Identifying Highly Correlated Features

To further examine the effect of key features on student's success rate, Multiple Linear Regression (MLR) was applied. This analysis identifies significant predictors of whether a student achieves 60% or higher. Based on the Kendall-Tau correlation analysis and prior regression modelling, the most statistically significant features were selected for MLR. The five strongest predictors identified were ATT (Attendance), GENI (Gender Inequality Faced by Students), HSS (Hours Spent on Self-Study), HCOC (Hours Spent in Coaching Classes), and HC (Hours Spent in Café/Canteen). These features were used as independent variables, while T (Target variable) was set as the dependent variable.

- **Multiple Linear Regression Equation**

The Multiple Linear Regression equation generated from the model is,

$$T = 0.1204(ATT) + 0.2086(\text{GENI}) + 0.0352(\text{HC}) + 0.0396(\text{HCOC}) + 0.2503(\text{HSS}) + \epsilon \quad (4.1)$$

where, T is the target variable, ATT is attendance, GENI is gender inequality faced by students, HC is hours spent in a café/canteen, HCOC is hours spent in coaching classes, and HSS is hours spent on self-study. The numeric values (0.1204, 0.2086, 0.0352, 0.0396, 0.2503) are the regression coefficients, indicating the strength of each feature's influence on T, while $\epsilon$ represents the error term. This equation demonstrates how each feature impacts a student's likelihood of achieving 60% or higher.

- **Regression Results and Interpretation**

The MLR results are detailed in Table 4.19, highlighting the regression coefficients, p-values, and significance levels.

**Table 4.19. Multiple Linear Regression Results**

| **Features** | **Coefficient** (β) | **p-value** | **Significance** |
|---|---|---|---|
| **ATT** | 0.1204 | 0.000 | Significant |
| **GENI** | 0.2086 | 0.000 | Significant |
| **HC** | 0.0352 | 0.050 | Borderline Significant |
| **HCOC** | 0.0396 | 0.001 | Significant |
| **HSS** | 0.2503 | 0.000 | Highly Significant |

The p-values indicate that all five features significantly contribute to predicting T ($p < 0.05$), confirming their crucial role in student academic success. Self-Study (HSS) has the highest positive impact ($\beta = 0.2503$, $p = 0.000$), proving that students who spend and utilize more time to self-study have a significantly higher probability of achieving 60% or higher. Gender Inequality (GENI) is also a strong predictor ($\beta = 0.2086$, $p = 0.000$), suggesting that students facing gender-related challenges put in additional effort to succeed. Attendance (ATT) positively influences ($\beta = 0.1204$, $p = 0.000$), implying that higher attendance is associated with better academic outcomes. Coaching Classes (HCOC, $\beta = 0.0396$, $p = 0.001$) and Café/ Canteen Time (HC, $\beta = 0.0352$, $p = 0.050$) also contribute positively, though with relatively smaller effects.

- **Model Performance**

The overall performance of the MLR model is summarized as, $R^2 = 0.073$ → The model explains 7.3% of the variance in student performance. F-statistic = 27.02 ($p < 0.0001$) → The model is statistically significant as a whole. No major multicollinearity issues were observed (Condition Number = 6.43), meaning the predictors are not excessively correlated. Although the $R^2$ value is relatively low, indicating that other factors may contribute to academic performance, the model confirms that the selected five features are statistically significant predictors.

### 4.1.4 Web-Based Interface: Results and Observations

This objective centers on developing an intuitive, web-based prediction system that utilizes the pre-trained ACO-DT model to assess student academic performance. Built with Django and Python, the system provides two functionalities, single-student prediction and multiple-student prediction, both accessible upon successful admin login. Figure 4.15 displays the login page, where users enter their information to use the system interface. Once logged in, the dashboard (Figure 4.16) appears, presenting various options through a menu that allows users to navigate to different functionalities. Figure 4.17 and 4.18 illustrates the data input page, where users can either enter student details manually or upload a dataset for bulk predictions. After submission, the prediction results page (Figure 4.19 and 4.20) displays classification outcomes, indicating whether a student is likely to score above or below 60% with recommendations.

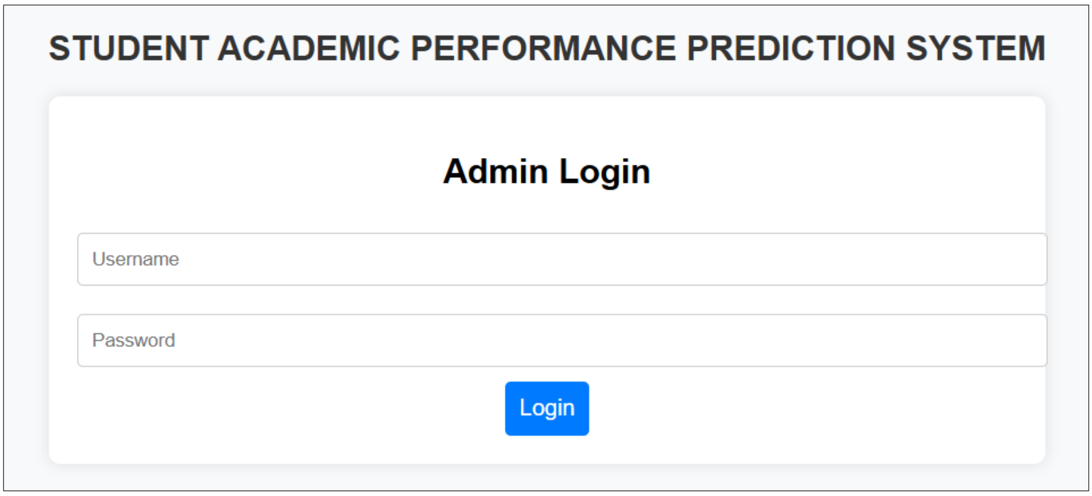


**Figure 4.15. Admin Login Page**

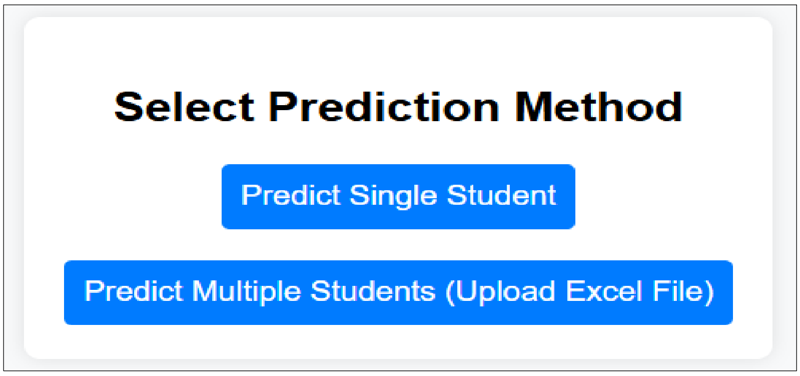


**Figure 4.16. Selection Dashboard**

### A. Scenario 1: Single Student prediction system

In this scenario, the user is presented with a form to enter data for an individual student. After submitting the required details, the system determines whether the student is likely to score above or below 60%. If the prediction indicates a score below 60%, personalized recommendations are provided to help improve academic performance. Additionally, the user has the option to generate a PDF report summarizing the prediction and recommendations. The interface is designed for simplicity, ensuring a seamless process from data input to result display. Figure 4.19 illustrates the result and recommendation page, highlighting the predicted outcomes along with tailored suggestions for students at risk.

### B. Scenario 2: Multiple Student prediction system

This scenario enables users to upload an Excel file containing data for multiple students. The system processes the file, predicts each student's academic performance, and generates personalized recommendations for those at risk of scoring below 60%. A PDF report summarizing all predictions and recommendations is then created. While the prediction and recommendation logic remain the same as in the single-student system, this scenario is designed for batch processing, allowing efficient handling of large datasets. The interface is optimized for bulk input and result display, making it particularly useful for educational institutions managing extensive student records. Figure 4.20 (a) and (b) illustrates the result and recommendation page for this interface.

### C. Scenario 3: Recommendation system

The web-based prediction system provides personalized recommendations to help students overcome various academic challenges. It encourages the development of effective study habits, goal setting, and time management by minimizing distractions such as excessive social media use, gaming, or extended time spent in non-productive spaces like canteens. For students balancing part-time jobs, the system suggests adjusting work hours to maintain a healthy work-study balance. To address financial difficulties, the system recommends scholarships, financial aid programs, and part-time job opportunities, while students from low-income backgrounds are encouraged to utilize institutional resources for academic support. It also emphasizes the importance of mentorship and counselling for students facing personal or family-related challenges, such as single-parent responsibilities, financial pressures, or large sibling families, ensuring they receive both emotional and practical guidance.

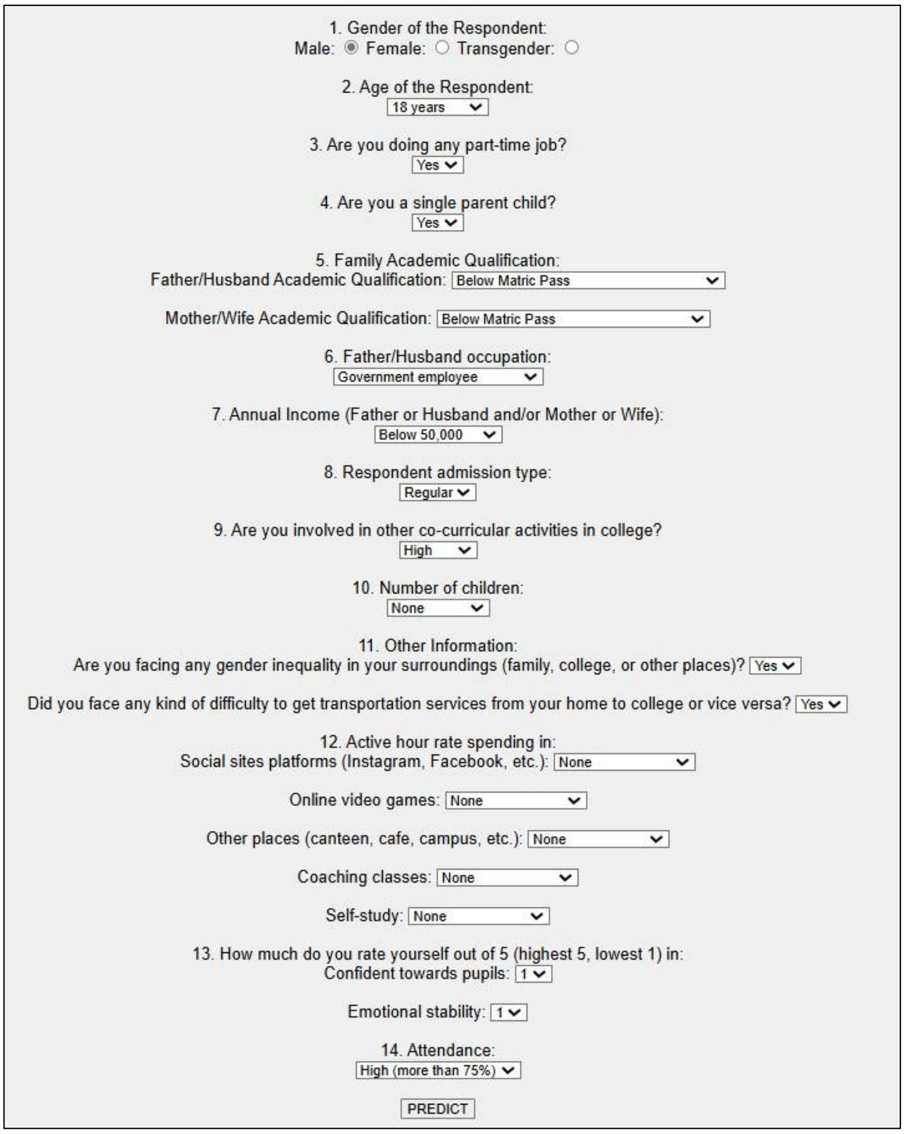


**Figure 4.17. Single Student Prediction Page**

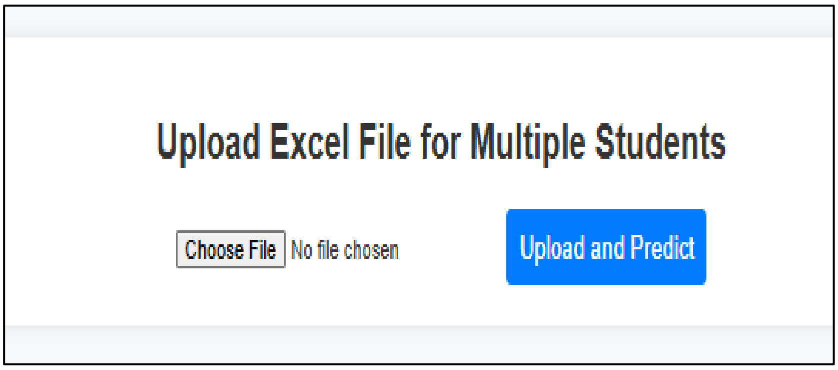


**Figure 4.18. Multiple Students Prediction Page**


**Figure 4.19. Result and Recommendation Page for Single-Students Prediction**

For those encountering societal barriers, such as gender inequality, the system promotes institutional support mechanisms, while LGBTQ+ students are encouraged to connect with inclusive campus groups for a supportive academic environment. Additionally, recommendations tailored to parental academic qualifications or family head occupations help students establish structured study routines, access community resources, and explore online learning opportunities. The system also identifies low attendance as a significant factor affecting academic performance, advising students to maintain consistent study habits and seek academic support to bridge knowledge gaps. By offering targeted and actionable insights, the system empowers students to overcome personal, social, and financial challenges, enabling them to focus on academic success and personal growth.

Upload Results

| GEN | AGE | PT_JOB | SIN_PAR | FAT_ACDQUA | MOT_ACDQUA | FH_OCCUP | TOT_INC | ADM_TYPE | CLG_CURRACT | N_CHILD | GEN_INQ | TRANS_DIFF | H_SOCSITE | H_VIDGAM | H_CANCAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 2 | 2 | 1 | 4 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| 0 | 2 | 1 | 0 | 3 | 3 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

Download PDF

**Figure 4.20 (a). Result and Recommendation Page for Multiple-Students Prediction**

Upload Results

| H_SOCSITE | H_VIDGAM | H_CANCAF | H_COCHCL | H_SELFST | R_CONF | R_EMOT | ATTEN | Prediction | Recommendations |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 2 | 4 | 3 | 1 | 0 | • Utilize campus resources and join study groups to address any gender-related barriers and boost your academic success.<br>• Balance your studies with work by prioritizing tasks and using campus resources to manage pressure.<br>• Consider reducing work hours or adjusting your schedule to better balance your studies and part-time job.<br>• Seek support from institute counselors or mentors to help manage responsibilities and stress.<br>• Create a structured study schedule and seek additional academic support to address any irregularities.<br>• Address any transport difficulty issues.<br>• Increase time and analysis your result.<br>• Practice mindfulness<br>• Improve your attendance |
| 2 | 0 | 0 | 0 | 2 | 4 | 5 | 1 | 0 | • Utilize campus resources and join study groups to address any gender-related barriers and boost your academic success.<br>• Maintain a healthy work-study balance and seek guidance from mentors to support your academic goals.<br>• Consider reducing work hours or adjusting your schedule to better balance your studies and part-time job.<br>• Seek support from institute counselors or mentors to help manage responsibilities and stress.<br>• Aim to establish a stable study routine and utilize community resources to support your academic needs.<br>• Explore scholarship opportunities and financial aid programs to support your academic needs and reduce financial stress.<br>• Increase time and analysis your result.<br>• Improve your attendance |
| 1 | 0 | 0 | 0 | 2 | 5 | 5 | 1 | 0 | • Utilize campus resources and join study groups to address any gender-related barriers and boost your academic success.<br>• Balance your studies with work by prioritizing tasks and using campus resources to manage pressure.<br>• Consider reducing work hours or adjusting your schedule to better balance your studies and part-time job.<br>• Seek support from institute counselors or mentors to help manage responsibilities and stress.<br>• Seek additional academic resources and support if needed to enhance your learning environment.<br>• Explore scholarship opportunities and financial aid programs to support your academic needs and reduce financial stress.<br>• Increase time and analysis your result. |

Download PDF

**Figure 4.20 (b). Result and Recommendation Page for Multiple-Students Prediction**

## 4.2. SUMMARY

This chapter detailed the findings and analysis of the student academic performance prediction system, covering feature selection, machine learning evaluation, optimization, feature analysis, and web-based implementation. The Chi-Square approach was chosen for further model assessment because it produced the highest accuracy (72.4%) among the nine feature selection strategies using logistic regression. Out of eight machine learning classifiers, the Decision Tree (DT) model achieved the highest accuracy (93.24%) when trained using SMOTE, 8-fold CV, and an 80:20 train-test split. With further optimization using the ABC and ACO algorithms, ACO-DT achieved an accuracy of 98.15%, surpassing ABC-DT. For ACO-DT with SMOTE, the optimal hyperparameters were, Max Depth=18, Min Samples Leaf=1, and Min Samples Split= 6. Furthermore, five significant predictors- attendance, gender inequality, self-study hours, coaching class hours, and café/canteen hours- were identified through Kendall-Tau correlation analysis and validated using Multiple Linear Regression (MLR). At last, a web-based system was developed using Django and Python. This system incorporates ACO-DT to provide real-time forecasts, personalized suggestions, and the ability to generate PDF reports. Academic achievement was strongly predicted by self-study, regular attendance, and coaching study support; however, social and economic difficulties were found to affect the performance. These findings may assist educators and legislators in implementing data-driven strategies to enhance student achievement.

# Chapter 5
# DISCUSSION AND CONCLUSION

This chapter discusses the key findings of the study, highlighting the effectiveness of feature selection, machine learning models, optimization techniques, and the web-based prediction system. The results are analyzed in relation to existing research, emphasizing the contributions of this study. Additionally, the chapter outlines the limitations of the current approach and suggests future directions for further research to enhance the predictive system.

## 5.1. DISCUSSION

### 5.1.1 Objective 1: To identify the most influential predictive variables affecting students' academic performance in higher education

To accomplish the first objective, an initial group of 42 state-of-the-art features was identified from existing literature and merged with 28 novel features proposed in this study, covering nine broad categories: demographic factors, educational factors, behavioral patterns, time-management skills, contextual influences, health factors, socioeconomic status, interpersonal factors, and family dynamics. After designing a comprehensive questionnaire and conducting a pilot study, data were collected, and nine feature selection techniques, Chi-Square, Pearson Correlation, Variance Threshold, RFE, Feature Importance, Random Forest, XGBoost, Lasso, and Ridge Regression were applied. Among these, the Chi-Square method delivered the best classification accuracy with the Logistic Regression model and was used to finalize 21 optimal features. A comparison of preliminary results before and after feature selection confirmed a significant performance improvement. Additionally, an analysis was performed to identify the relevant features suggested by each technique, followed by a detailed relevance analysis using the Chi-Square method.

To further identify the significance of the selected features, two statistical techniques, Kendall Tau correlation and Multiple Linear Regression (MLR) were applied. Kendall Tau, a statistical technique, suitable for ranking-style data, is useful in checking which features seemed to associate most with students scoring 60% or higher. Interestingly, five features stood out: attendance, gender-based challenges, confidence levels, the type of admission, and family income. The values weren't super high but still meaningful, 0.28 for attendance, then 0.236, 0.162, 0.154, and 0.107 for the rest. So, based on this, it proves, students who regularly attend classes, feel confident, face fewer gender-related issues, come through regular admissions, and come from more financially stable families tend to do better.

In addition to Kendall Tau, the combined influence of crucial factors on academic achievement was determined using Multiple Linear Regression (MLR). Attendance, Gender Inequality, Café/Canteen Hours, Coaching Class Hours, and Self-Study Hours were the five independent features used in this research, with the target variable (T) serving as the dependent feature. The regression results showed that all five features significantly contributed to predicting academic outcomes, with Self-Study having the strongest impact, followed by Gender Inequality, Attendance, Coaching Classes, and Café/ Canteen Time. While Kendall Tau focuses on individual relationships, MLR accounts for the combined predictive effect, which explains the variation in the top features identified by each method. The common presence of Attendance and Gender Inequality in both analyses reinforces their importance in influencing student academic performance.

The results of this objective are detailed in Chapter 4, Sections 4.1.1 and 4.1.4.

### 5.1.2 Objective 2: To formulate machine learning predictive model of selected parameters

To fulfill the second objective, various machine learning algorithms were explored to formulate a predictive model using the selected optimal features. Eight widely used classifiers, Decision Tree, Naïve Bayes, Support Vector Machine, Random Forest, XGBoost, and others were evaluated under multiple experimental setups. These included the use of SMOTE to handle the issue of class imbalance, cross-validation with 8-fold and 10-fold strategies to ensure generalizability, and a traditional 80:20

train-test split for performance validation. Among all models, the Decision Tree classifier emerged as the best-performing algorithm. Specifically, when trained using SMOTE with 8-fold cross-validation, the DT classifier achieved the highest accuracy of 93.24%, along with an exceptional precision of 98.93%, recall of 96.69%, and F1-score of 97.80%. These metrics prove that the model not only correctly classified most of the instances but also maintained consistency across performance measures. The use of SMOTE significantly enhanced the classifier's capability to handle imbalanced data, which is a common challenge in educational datasets. Thus, the Decision Tree model was chosen as the best predictive model for determining whether a student is likely to succeed academically (defined as scoring 60% or more) after a thorough comparative examination.

To further enhance model accuracy, the Decision Tree model was optimized using two metaheuristic algorithms, namely ABC and ACO. These methods were employed to fine-tune the DT classifier's hyperparameters- Min Samples Split, Min Samples Leaf and Max Depth. The ACO-DT variant surpassed other ABC-DT model by achieving a 98.15% classification accuracy. When SMOTE was used in conjunction with hyperparameter tuning, it contributed to improved model performance. The optimal and finest parameters consisted of Max Depth of 18, Min Samples Leaf of 1, and Min Samples Split of 6. The research findings demonstrate that such optimization algorithms are effective in enhancing machine learning algorithms, thereby improving the predictive accuracy of student academic performance.

The results of this objective are detailed in Chapter 4, Sections 4.1.2 and 4.1.3.

### 5.1.3 Objective 3: To develop a web based predictive model interface

A web-based application implemented the prediction model for real-time application as one of the research's main objectives. Developed in Python and the Django framework within the PyCharm IDE, this interface provides a dynamic platform for educational institutions to provide feature-based performance predictions for their students. To accommodate a variety of user requirements, the interface can be configured to receive inputs either manually (for forecasts involving a single student) or by uploading an Excel file (for predictions involving numerous students).

The ACO-optimized Decision Tree model is integrated into the interface, and the 21 optimum features chosen using the Chi-Square approach are used for prediction. Users may get quick results by clicking the "Predict" button after entering the necessary data. The result makes it very evident whether the trained model predicts a student will get a score of 60% or above. The interface also provides personalised suggestions to help students who are expected to get a score lower than 60%.

In the case of multiple student predictions, the system processes the uploaded Excel file, displays the results on a separate result page, and provides the option to generate a PDF report. This report is automatically downloaded, offering a convenient summary of academic predictions for administrative or academic review. The interface was evaluated using datasets from five separate institutions, proving its adaptability and practicality in different educational settings.

Overall, the deployment of this web-based tool translates the theoretical model into a practical application, enabling timely interventions, early identification of underperforming students, and improved decision-making process for educators and institutions.

The results of this objective are detailed in Chapter 4, Section 4.1.5.

### 5.1.4 Objective 4: To test and validate the proposed model with state-of-the-art models

To address the last objective, the performance of the proposed ACO-optimized Decision Tree (ACO-DT) model was rigorously tested and validated against a range of traditional ML classifiers and findings from research literature. All models were trained and evaluated under consistent experimental settings, including SMOTE-based class balancing, 8-fold CV, and an 80:20 train-test split, to ensure fairness and reliability in the comparison.

Among all the models, the ACO-DT model demonstrated superior performance, achieving an accuracy of 98.15%. The use of Ant Colony Optimization contributed to enhanced feature exploration and better decision boundaries, significantly improving the model's generalization ability.

In contrast, traditional classifiers such as Naïve Bayes and basic SVM exhibited lower predictive accuracy, particularly in imbalanced data scenarios, despite the application of SMOTE. While XGBoost and Random Forest offered competitive performance, they were still outperformed by the ACO-DT model in both accuracy and computational efficiency.

This comprehensive evaluation confirms that the integration of metaheuristic optimization, particularly ACO, with a Decision Tree classifier yields a more robust and accurate predictive model. The results highlight the effectiveness of the proposed model in outperforming well-established traditional models and establish its potential for practical deployment in academic environments to support data-driven decision-making.

The results of this objective are detailed in Chapter 4, Section 4.1.3 (C).

## 5.2. CONCLUSION

In summary, the study achieved all four research objectives effectively. From Objective 1, twenty-one optimal features were finalized for predicting students' academic performance and campus placement eligibility, with Self-Study, Gender Inequality, Attendance, Coaching Classes, and Café Hours identified as the most influential factors. Objective 2 involved developing and evaluating eight machine learning models, where the Decision Tree classifier achieved the highest performance (93.24% accuracy) and further improved to 98.15% after optimization using the Ant Colony Optimization (ACO) technique. Objective 3 focused on designing a web-based prediction system using Python and Django that integrates the ACO-optimized Decision Tree model with the finalized 21 features, enabling real-time single and multiple student predictions with downloadable reports. Finally, Objective 4 validated the proposed ACO-DT model against state-of-the-art approaches, confirming its superior accuracy (98.15%) and computational efficiency, demonstrating the effectiveness of integrating machine learning with metaheuristic optimization for academic performance prediction.

Additionally, beyond student-focused implications, the study's findings hold significance for institutions, policymakers, and curriculum designers. The ACO-optimized Decision Tree model and web-based system can assist in early identification of at-risk students and institutional performance monitoring. Such predictive insights can guide evidence-based interventions, curriculum improvements, and policy decisions to enhance the overall quality of higher education.

## 5.3. LIMITATIONS OF THE STUDY

While the present research successfully developed a robust and accurate prediction model, certain aspects offer valuable opportunities for further enhancement. The dataset, though comprehensive within the Computer Science domain, can be enriched by including students from diverse academic streams and institutions to broaden its applicability. Although the web-based system has been successfully designed and tested, its large-scale deployment and real-time use in institutional settings are yet to be implemented. Furthermore, while the questionnaire effectively captured significant academic, social, and demographic parameters, it can be periodically updated to align with evolving educational needs and technological advancements. Addressing these aspects in future work will enhance the model's practical applicability and ensure its continued relevance in dynamic learning environments.

## 5.4. FUTURE WORK

**5.4.1. Expanding the dataset:** Future research can focus on enlarging the dataset by including students from various academic disciplines, colleges, and geographic regions. This will help to validate the model across diverse educational settings, ensuring its scalability and robustness. By integrating multidisciplinary data, the system could better capture differences in learning patterns, academic challenges, and institutional factors influencing performance.

**5.4.2. Integration of advanced deep learning models:** Although machine learning techniques have shown effective predictive capabilities, incorporating deep learning architectures such as, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks can further improve performance. CNNs can capture complex nonlinear relationships among academic and behavioral variables, while LSTMs are particularly useful for identifying time-dependent patterns, such as changes in student performance across semesters or academic years.

**5.4.3. Real-world deployment and institutional integration:** A key future direction is the deployment of the developed web-based prediction system as a publicly accessible platform. Its integration with institutional and national educational quality frameworks (such as NAAC) can transform it into a real-time academic monitoring and decision-support tool. This would enable teachers, administrators, and policymakers to proactively identify at-risk students and provide timely interventions, fostering an evidence-based approach to academic improvement.

**5.4.4. Longitudinal analysis and feedback-driven improvement:** Future studies should also include longitudinal tracking of students to compare predicted outcomes with actual performance over time. This would provide valuable feedback for continuous model refinement and validation. By integrating real-time performance updates, the prediction system could evolve into a dynamic, self-learning tool that adapts to changing academic and behavioral trends.

# REFERENCES


[1] I. Mushtaq and S. Nawaz Khan, "Factors Affecting Students' Academic Performance," *Global Journal of Management and Business Research*, vol. 12, no. 9, 2012, [Online]. Available: https://www.academia.edu/Jobs%0Ahttps://www.academia.edu/hiring

[2] M. Hussain, W. Zhu, W. Zhang, S. Muhammad, R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning session data," *Artif Intell Rev*, 2018, doi: 10.1007/s10462-018-9620-8.

[3] A. O. Alsayed *et al.*, "Selection of the Right Undergraduate Major by Students Using Supervised Learning Techniques," *applied sciences*, vol. 11, 2021.

[4] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student academic performance prediction using supervised learning techniques," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, pp. 92–104, 2019, doi: 10.3991/ijet.v14i14.10310.

[5] K. Mahawar and P. Rattan, "A Comprehensive Investigation into the Dimensions of Educational Data Mining Using Artificial Intelligence," in *International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, IEEE, 2023. Accessed: Jun. 02, 2024. [Online]. Available: 10.1109/ICACCTech61146.2023.00014

[6] Y. K. Saheed, T. O. Oladele, A. O. Akanni, and W. M. Ibrahim, "Student performance prediction based on data mining classification techniques," *Nigerian Journal of Technology*, vol. 37, no. 4, p. 1087, Nov. 2018, doi: 10.4314/njt.v37i4.31.

[7] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Educ Inf Technol (Dordr)*, vol. 26, no. 1, pp. 205–240, 2021, doi: 10.1007/s10639-020-10230-3.

[8] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[9] L. Khanna, S. N. Singh, and M. Alam, "Educational data mining and its role in determining factors affecting students academic performance: A systematic review," *India International Conference on Information Processing, IICIP 2016 - Proceedings*, 2016, doi: 10.1109/IICIP.2016.7975354.

[10] Q. Nguyen and A. Arbor, "Rethinking time-on-task estimation with outlier detection accounting for individual , time , and task differences," *In Proceedings of the International Conference on Learning Analytics and Knowledge, Frankfurt, Germany*, pp. 376–381, 2020, doi: 10.1145/3375462.3375538.

[11] S. Kumar, Dr. M. Agarwal, and Dr. N. Agarwal, "Defining and measuring academic performance of HEI students-A critical review," *Turkish Journal of Computer and Mathematics Education*, vol. 3091, no. 6, pp. 3091–3105, 2021, [Online]. Available:

https://www.researchgate.net/figure/Astins-Theory-of-Involvement-Inputs-Environments-

[12] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, 2020, doi: 10.1186/s41239-020-0177-7.

[13] B. Adeyeye *et al.*, "Online Learning Platforms and Covenant University Students' Academic Performance in Practical Related Courses during COVID-19 Pandemic," *Sustainability (Switzerland)*, vol. 14, no. 2, 2022, doi: 10.3390/su14020878.

[14] B. K. Francis and S. S. Babu, "Predicting Academic Performance of Students Using a Hybrid Data Mining Approach," *J Med Syst*, vol. 43, no. 6, 2019, doi: 10.1007/s10916-019-1295-4.

[15] K. Mahawar and P. Rattan, "Dimensionality Reduction using Feature Selection Techniques on EDM for Student Academic Performance Prediction," 2023. doi: https://doi.org/10.17762/ijritcc.v11i10.8961.

[16] B. K. Yousafzai, M. Hayat, and S. Afzal, "Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student," *Educ Inf Technol (Dordr)*, 2020.

[17] E. Mahsereci, S. Ay, and İ. Turgay, "A comparative study on the effect of feature selection on classification accuracy," *Procedia Technology*, vol. 1, pp. 323–327, 2012, doi: 10.1016/j.protcy.2012.02.068.

[18] M. Zaffar, M. A. Hashmani, K. S. Savita, S. S. H. Rizvi, and M. Rehman, "Role of FCBF Feature Selection in Educational Data Mining," *Mehran University Research Journal of Engineering and Technology*, vol. 39, no. 4, pp. 772–778, 2020, doi: 10.22581/muet1982.2004.09.

[19] Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance," *Comput Intell Neurosci*, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/4151487.

[20] M. Doshi and S. K. Chaturvedi, "Correlation Based Feature Selection (CFS) Technique to Predict Student Perfromance," *International journal of Computer Networks & Communications*, vol. 6, no. 3, pp. 197–206, 2014, doi: 10.5121/ijcnc.2014.6315.

[21] M. Zaffar, M. A. Hashmani, K. S. Savita, and S. S. H. Rizvi, "A study of feature selection algorithms for predicting students academic performance," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, pp. 541–549, 2018, doi: 10.14569/IJACSA.2018.090569.

[22] M. A. Naaj, R. Mehdi, and E. A. Mohamed, "Analysis of the Factors Affecting Student Performance Using a Neuro-Fuzzy Approach," *Educ Sci (Basel)*, 2023.

[23] M. Zaffar *et al.*, “A hybrid feature selection framework for predicting students performance,” *Computers, Materials and Continua*, vol. 70, no. 1, pp. 1893–1920, 2021, doi: 10.32604/cmc.2022.018295.

[24] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, “Predicting Students Performance in Educational Data Mining,” *Proceedings - 2015 International Symposium on Educational Technology, ISET 2015*, pp. 125–128, 2015, doi: 10.1109/ISET.2015.33.

[25] J. Macleod, H. H. Yang, S. Zhu, and Y. Shi, “Technological Factors and Student-to-Student Connected Classroom Climate in Cloud Classrooms,” *Journal of Educatonal Computing Research*, 2017, doi: 10.1177/0735633117733999.

[26] N. Robert *et al.*, “Determining factors that affect student performance using various machine learning methods machine learning methods,” *Procedia Comput Sci*, vol. 216, no. 2022, pp. 597–603, 2023, doi: 10.1016/j.procs.2022.12.174.

[27] M. E. Işikgöz, “Analysis of the Relationship between Online Learning Activities and Academic Achievement of Physical Education and Sports School Students,” *The Turkish Online Journal of Educational Technology*, vol. 22, no. 1, 2023.

[28] I. D. Juguilon, “Impact of Family Support System in the Academic Performance of Grade 3 Pupils at a Public Elementary School in Rizal, Philippines,” *INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY: APPLIED BUSINESS AND EDUCATION RESEARCH*, vol. 4, no. 1, pp. 174–187, 2023, doi: 10.11594/ijmaber.04.01.16.

[29] R. Saxena, “Educational Data Mining : Performance Evaluation of Decision Tree and Clustering Techniques Using WEKA Platform,” *International Journal of Computer Science and Business Informatics*, vol. 15, no. 2, 2015.

[30] H. Yang, J. Cai, H. Hao, and X. Wang, “Examining key factors of beginner ’ s continuance intention in blended learning in higher education,” *J Comput High Educ*, vol. 35, no. 1, pp. 126–143, 2023, doi: 10.1007/s12528-022-09322-5.

[31] K. T. Chui, D. Chun, L. Fung, M. D. Lytras, and T. M. Lam, “Predicting At-risk University Students in a Virtual Learning Environment via a Machine Learning Algorithm,” *Comput Human Behav*, 2018, doi: 10.1016/j.chb.2018.06.032.

[32] A. N. Maestre, M. A. vega Rodriguez, C. J. Perez, F. C. Alonso, and J. M. S. Gomez, “Social Network Analysis and Educational Data Mining for NeuroK e-Learning Platform,” *Creat Educ*, no. 1, pp. 23–28, 2017, [Online]. Available: https://www.academia.edu/download/54164684/ICETIC_FullPapers.pdf#page=31

[33] J. A. Paul, H. M. Baker, and J. D. Cochran, “Effect of online social networking on student academic performance,” *Comput Human Behav*, vol. 28, no. 6, pp. 2117–2127, 2012, doi: 10.1016/j.chb.2012.06.016.

[34] L. Cheng, K. R. Varshney, and H. Liu, “Socially Responsible AI Algorithms: Issues, Purposes, and Challenges,” 2021.

[35] E. T. Lau, L. Sun, and Q. Yang, "Modelling, prediction and classification of student academic performance using artificial neural networks," *SN Appl Sci*, vol. 1, no. 9, pp. 1–10, 2019, doi: 10.1007/s42452-019-0884-7.

[36] V. Dhanalakshmi, D. Bino, and S. A. M., "Opinion mining from student feedback data using supervised learning algorithms," *IEEE, 3rd MEC International Conference on Big Data and Smart City*, pp. 1–5, 2016.

[37] L. Zhang and M. Komachi, "Chinese – Japanese Unsupervised Neural Machine Translation Using Sub-character Level Information," *Computation and Language*, 2018.

[38] K. Mahawar and P. Rattan, "Empowering education: Harnessing ensemble machine learning approach and ACO-DT classifier for early student academic performance prediction," *Educ Inf Technol (Dordr)*, Sep. 2024, doi: 10.1007/s10639-024-12976-6.

[39] Brett. Lantz, *Machine learning with R*. Packt Publishing, 2013.

[40] C. Kaestner, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

[41] Gutierrez Daniel D., *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R*. 2015.

[42] D. Chahal and L. Kharb, "Smart Diagnosis of Orthopaedic Disorders using Internet of Things (IoT)," *Int J Eng Adv Technol*, no. 2249–8958, 2019, [Online]. Available: http://www.cbsr.ia.ac.cn/irisdatabase.html.

[43] K. Deepika and N. Sathyanarayana, "Relief-F and budget tree random forest based feature selection for student academic performance prediction," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, pp. 30–39, 2019, doi: 10.22266/IJIES2019.0228.04.

[44] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interactive Learning Environments*, vol. 0, no. 0, pp. 1–20, 2021, doi: 10.1080/10494820.2021.1928235.

[45] C. Zabriskie, J. Yang, S. Devore, and J. Stewart, "Using machine learning to predict physics course outcomes," *Phys Rev Phys Educ Res*, vol. 15, no. 2, p. 20120, 2019, doi: 10.1103/PhysRevPhysEducRes.15.020120.

[46] J. Y. Chung and S. Lee, "Dropout early warning systems for high school students using machine learning," *Child Youth Serv Rev*, pp. 1–35, 2018, doi: 10.1016/j.childyouth.2018.11.030.

[47] H. Waheed *et al.*, "Balancing sequential data to predict students at-risk using adversarial networks," *Computers and Electrical Engineering*, vol. 93, no. September 2020, p. 107274, 2021, doi: 10.1016/j.compeleceng.2021.107274.

[48] S. Gavankar and S. Sawarkar, "Decision tree: Review of techniques for missing values at training, testing and compatibility," *Proceedings - AIMS 2015, 3rd International Conference on Artificial Intelligence, Modelling and Simulation*, pp. 122–126, 2015, doi: 10.1109/AIMS.2015.29.

[49] M. Wafi, U. Faruq, and A. A. Supianto, "Automatic Feature Selection for Modified K-Nearest Neighbor to Predict Student's Academic Performance," *Proceedings of 2019 4th International Conference on Sustainable Information Engineering and Technology, SIET 2019*, pp. 44–48, 2019, doi: 10.1109/SIET48054.2019.8986074.

[50] A. Shankhdhar, D. Sharma, M. Pushkarna, A. Agrawal, and S. Chaturvedi, "Intelligent Decision Support System Using Decision Tree Method for Student Career," *International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, pp. 140–142, 2020.

[51] Y. Nieto, V. G. Carlos, M. Rubén, and G. Crespo, "Supporting academic decision making at higher educational institutions using machine learning-based algorithms," *Soft comput*, 2018, doi: 10.1007/s00500-018-3064-6.

[52] M. Ciolacu, A. F. Tehrani, L. Binder, and P. M. Svasta, "Education 4 . 0 - Artificial Intelligence assisted Higher Education : Early recognition System with Machine Learning to support Students ' Success," *2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, no. Ml, pp. 23–30, 2018.

[53] V. Shaga, H. Gebregziabher, and P. Chintal, "Predicting Performance of Students Considering Individual Feedback at Online Learning Using Logistic Regression Model," *Lecture Notes in Networks and Systems*, vol. 191, no. Ictcs 2020, pp. 111–120, 2022, doi: 10.1007/978-981-16-0739-4_11.

[54] L. Wang, "Metaheruistic optimization based machine-learning approach for analysis of academic students' strategies," *Journal of Intelligent & Fuzzy Systems*, pp. 1–13, Apr. 2024, doi: 10.3233/JIFS-234021.

[55] X. Xu, Y. Lei, and Z. Li, "An Incorrect Data Detection Method for Big Data Cleaning of Machinery Condition Monitoring," *IEEE Transactions on Industrial Electronics*, vol. PP, no. c, p. 1, 2019, doi: 10.1109/TIE.2019.2903774.

[56] F. Tan, "Improving Feature Selection Techniques for machine learning," *Dimension Contemporary German Arts And Letters*, p. 110, 2007, [Online]. Available: http://etd.gsu.edu/theses/available/etd-11262007-015841/unrestricted/tan_feng_200712_phd.pdf

[57] I. Brajevic, T. Milan, and M. Subotic, "Performance of the improved artificial bee colony algorithm on standard engineering constrained problems," *International Journal of Mathematics and Computers in Simulation*, vol. 5, no. 2, 2011.

[58] N. Karaboğa and M. B. Çetinkaya, "A novel and efficient algorithm for adaptive filtering: Artificial bee colony algorithm," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 19, no. 1, pp. 175–190, 2011, doi: 10.3906/elk-0912-344.

[59] D. Teodorović and M. Dell' Orco, "BEE COLONY OPTIMIZATION-A COOPERATIVE LEARNING APPROACH TO COMPLEX TRANSPORTATION PROBLEMS," *Advanced OR and AI Methods in Transportation*, pp. 51–60, 2003.

[60] Y. Zhang, "Application of data mining based on Improved Ant Colony Algorithm in college students' employment and Entrepreneurship Education," *Res Sq*, 2023, doi: 10.21203/rs.3.rs-2769200/v1.

[61] J. Sobecki and J. M. Tomczak, "Student Courses Recommendation Using Ant Colony Optimization," *Intelligent Information and Database Systems*, vol. 5991, pp. 124–133, 2010.

[62] J. Zhao, H. Mao, P. Mao, and J. Hao, "Learning path planning methods based on learning path variability and ant colony optimization," *Systems and Soft Computing*, vol. 6, Dec. 2024, doi: 10.1016/j.sasc.2024.200091.

[63] H. Hassan, N. B. Ahmad, and S. Anuar, "Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining," *J Phys Conf Ser*, vol. 1529, no. 5, 2020, doi: 10.1088/1742-6596/1529/5/052041.

[64] A. Abid, I. Kallel, J. J. Sanchez-Medina, and M. Ben Ayed, "Parameters Sensitivity Analysis of Ant Colony Based Clustering: Application for Student Grouping in Collaborative Learning Environment," *IEEE Access*, vol. 12, pp. 24751–24761, 2024, doi: 10.1109/ACCESS.2023.3279723.

[65] A. Dirin and C. A. Saballe, "Machine Learning Models to Predict Students' Study Path Selection," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 1, pp. 158–183, 2022, doi: 10.3991/IJIM.V16I01.20121.

[66] G. Akçapınar, A. Altun, and A. Petek, "Using learning analytics to develop early-warning system for at-risk students," *International Journal of Educational Technology in Higher Education*, pp. 1–20, 2019.

[67] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J Bus Res*, vol. 94, no. August 2017, pp. 335–343, 2019, doi: 10.1016/j.jbusres.2018.02.012.

[68] T. Nazaretsky, M. Cukurova, M. Ariely, and G. Alexandron, "Confirmation bias and trust : Human factors that influence teachers ' attitudes towards AI-based educational technology," *CEURWorkshop Proceedings*, 2021.

[69] I. Katib and M. Ragab, “Harnessing variable reduction approach with deep recurrent neural network for student’s academic performance analysis,” *Alexandria Engineering Journal*, vol. 118, pp. 393–405, Apr. 2025, doi: 10.1016/j.aej.2025.01.082.

[70] A. Dutt, M. A. Ismail, and T. Herawan, “A Systematic Review on Educational Data Mining,” *IEEE, Systematic Review on Educational Data Mining*, vol. 5, pp. 15991–16005, 2017.

[71] J. A. Ali, A. H. Muse, M. K. Abdi, T. A. Ali, Y. H. Muse, and M. A. Cumar, “Machine learning-driven analysis of academic performance determinants: Geographic, socio-demographic, and subject-specific influences in Somaliland’s 2022–2023 national primary examinations,” *International Journal of Educational Research Open*, vol. 8, Jun. 2025, doi: 10.1016/j.ijedro.2024.100426.

[72] C. Hsu, C. Chun, Y. Huang, and Y. Yu, “Effects of augmented reality - based multidimensional concept maps on students ’ learning achievement , motivation and acceptance,” *Univers Access Inf Soc*, no. 1, 2017, doi: 10.1007/s10209-017-0595-z.

[73] Á. H. Galvis, “Supporting decision-making processes on blended learning in higher education : literature and good practices review,” *International Journal of Educational Technology in Higher Education*, pp. 1–38, 2018.

[74] S. O. da Fonseca and A. A. Namen, “DATA MINING ON INEP DATABASES : AN INITIAL ANALYSIS AIMING TO IMPROVE BRAZILIAN EDUCATIONAL SYSTEM,” *Educação em Revista*, vol. 32, no. 01, pp. 133–157, 2016.

[75] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Multi-split optimized bagging ensemble model selection for multi-class educational data mining,” *Applied Intelligence*, vol. 50, no. 12, pp. 4506–4528, Dec. 2020, doi: 10.1007/s10489-020-01776-3.

[76] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, “Educational data mining and analysis of students’ academic performance using WEKA,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 2, pp. 447–459, 2018, doi: 10.11591/ijeecs.v9.i2.pp447-459.

[77] E. Sugiyarti, K. A. Jasmi, B. Basiron, and M. Huda, “DECISION SUPPORT SYSTEM OF SCHOLARSHIP GRANTEE SELECTION USING DATA MINING,” *International Journal of Pure and Applied Mathematics*, vol. 119, no. 15, pp. 2239–2249, 2018.

[78] S. Helal *et al.*, “Identifying key factors of student academic performance by subgroup,” *Int J Data Sci Anal*, vol. 7, no. April, pp. 227–245, 2020.

[79] C. Burgos, M. L. Campanario, D. De, J. A. Lara, D. Lizcano, and M. A. Martínez, “Data mining for modeling students’ performance: A tutoring action plan to prevent

academic dropout," *Computers and Electrical Engineering*, vol. 0, pp. 1–16, 2017, doi: 10.1016/j.compeleceng.2017.03.005.

[80] C. Vialardi, J. Bravo, L. Shafti, and A. Ortigosa, "Recommendation in Higher Education Using Data Mining Techniques," *Educational Data Mining*, 2009.

[81] C. Obeid, I. Lahoud, H. El Khoury, and P.-A. Chamin, "Ontology-based Recommender System in Higher Education," *The Third Edition of Educational Knowledge Management Workshop*, vol. 2, pp. 1031–1034, 2018.

[82] F. Ricci, L. Rokach, and B. Shapira, "Recommender Systems : Introduction and Challenges," in *Recommender Systems Handbook*, 2015, pp. 1–34. doi: 10.1007/978-1-4899-7637-6.

[83] J. Ren, J. Guo, W. Qian, X. Hao, and H. Jingjing, "Building an Effective Intrusion Detection System by Using Hybrid Data Optimization Based on Machine Learning Algorithms," *Security and Communication Networks*, 2019.

[84] X. Li, L. Goa, Q. Pan, L. Wan, and K. Chao, "An Effective Hybrid Genetic Algorithm and Variable Neighborhood Search for Integrated Process Planning and Scheduling in a Packaging Machine Workshop," *IEEE Trans Syst Man Cybern Syst*, 2018, doi: 10.1109/TSMC.2018.2881686.

[85] A. Aziz, R. Abou-samra, and A. Aprilianto, "Online Assessment of Islamic Religious Education Learning," *Tafkir: Interdisciplinary Journal of Islamic Education*, vol. 3, no. 1, pp. 60–76, 2022.

[86] N. Pandey, D. Kumar, A. Hudait, and A. Sen, "Automated classification of software issue reports using machine learning techniques : an empirical study," *Innov Syst Softw Eng*, 2017, doi: 10.1007/s11334-017-0294-1.

[87] M. Deschênes, "Recommender systems to support learners ' Agency in a Learning Context : a systematic review," *International Journal of Educational Technology in Higher Education*, pp. 1–23, 2020.

[88] J. Du, K. Foon, and T. Hew, "Using recommender systems to promote self- regulated learning in online education settings : current knowledge gaps and suggestions for future research," *Journal of Research on Technology in Education*, vol. 0, no. 0, pp. 1–22, 2021, doi: 10.1080/15391523.2021.1897905.

[89] E. Murgia, M. Landoni, T. Huibers, J. A. Fails, and M. S. Pera, "The Seven Layers of Complexity of Recommender Systems for Children in Educational Contexts," *CEUR Workshop Proc*, pp. 5–9, 2019.

[90] H. Aldowah, H. Al-samarraie, and W. M. Fauzy, "Educational Data Mining and Learning Analytics for 21stcentury higher education: A Review and Synthesis," *Telematics and Informatics*, 2019, doi: 10.1016/j.tele.2019.01.007.

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

[92] L. F. Daghestani, L. F. Ibrahim, R. S. Al Towirgi, and H. A. Salman, "Adapting gamified learning systems using educational data mining techniques," *Computer Applied Engineering Education*, pp. 1–22, 2020, doi: 10.1002/cae.22227.

[93] L. G. Camacho-Morles, J., Slemp, G. R., Pekrun, R., Loderer, K., Hou, H. and Oades, "Activity achievement emotions and academic performance: A meta-analysis. Manuscript under review.," *Educ Psychol Rev*, 2020.

[94] M. A. Job, "Data Mining Techniques Applying on Educational Dataset to Evaluate Learner Performance Using Cluster Analysis," *European Journal of Engineering Research and Science*, vol. 3, no. 11, pp. 25–31, 2018.

[95] B. Alharbi, F. Assiri, and B. Alharbi, "A Comparative Study of Student Performance Prediction using Pre-Course Data," *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 3, pp. 1–14, 2015.

[96] A. M. Jiménez-carvelo, A. González-casado, M. Gracia, and L. Cuadros-rodríguez, *Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review*. Elsevier Ltd, 2019. doi: 10.1016/j.foodres.2019.03.063.

[97] D. Buenaño-fernandez, S. Luján-mora, M. González, and D. Gil, "Text Mining of Open-Ended Questions in Self-Assessment of University Teachers : An LDA Topic Modeling Approach," *IEEE Access, SPECIAL SECTION ON ADVANCED DATA MINING METHODS FOR SOCIAL COMPUTING*, vol. 8, pp. 35318–35330, 2020, doi: 10.1109/ACCESS.2020.2974983.

[98] A. Onan, "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach," *Computer Applications in Engineering Education*, vol. 29, no. 3, pp. 572–589, 2021, doi: 10.1002/cae.22253.

[99] J. Kay and B. Kummerfeld, "From data to personal user models for life-long , life-wide learners," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 2871–2884, 2019, doi: 10.1111/bjet.12878.

[100] F. Bouchet, J. M. Harley, G. J. Trevors, and R. Azevedo, "Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning," *JEDM - Journal of Educational Data Mining*, vol. 5, no. 1, pp. 104–146, 2013, [Online]. Available: http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/32

[101] S. Neslihan, A. N. Gorban, J. Levesley, and E. M. Mirkes, "Automatic short answer grading and feedback using text mining methods," *Procedia Comput Sci*, vol. 169, no. 2019, pp. 726–743, 2020, doi: 10.1016/j.procs.2020.02.171.

[102] X. Xie, Y. Fu, H. Jin, Y. Zhao, W. Cao, and Y. Fu, "A novel text mining approach for scholar information extraction from web content in Chinese," *Future Generation Computer Systems*, 2019, doi: 10.1016/j.future.2019.08.033.

[103] C. Tseng, "Text Mining Analysis of Teaching Evaluation Questionnaires for the Selection of Outstanding Teaching Faculty Members," *IEEE Access*, vol. 6, pp. 72870–72879, 2018, doi: 10.1109/ACCESS.2018.2878478.

[104] N. Spatiotis, I. Mporas, I. Perikos, and M. Paraskevas, "Evaluation of an Educational Training Platform Using Text Mining," *Association for Computing Machinery*, pp. 10–14, 2018.

[105] R. F. Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *Data Mining Knowlede Discovery*, 2019, doi: 10.1002/widm.1332.

[106] P. K. Shaalan, S. Salloum, and A. Mohamed, "Using Text Mining Techniques for Extracting Information from Research Articles," *Intelligent Natural Language Processing: Trends and Applications*, 2018.

[107] H. Ghorbani, "MAHALANOBIS DISTANCE AND ITS APPLICATION FOR DETECTING MULTIVARIATE OUTLIERS," *SERIES MATHEMATICS AND INFORMATICS*, vol. 34, no. 3, pp. 583–595, 2019.

[108] Dr. A. T. Y. A. A. Alazeez, "ODED: Outlier Detection in Educational Data," *Iraqi Journal of Statistical Science*, no. 33, pp. 72–88, 2021.

[109] M. Aslam, "On detecting outliers in complex data using Dixon ' s test under neutrosophic statistics," *J King Saud Univ Sci*, vol. 32, no. 3, pp. 2005–2008, 2020, doi: 10.1016/j.jksus.2020.02.003.

[110] T. Dharmawan, H. Ginardi, and A. Munif, "Dropout Detection Using Non-Academic Data," *4TH Conference on Science and Technology*, 2018.

[111] J. Li, J. Zhang, N. Pang, and X. Qin, "Weighted Outlier Detection of High-Dimensional Categorical Data Using Feature Grouping," *IEEE Trans Syst Man Cybern Syst*, vol. PP, pp. 1–14, 2018, doi: 10.1109/TSMC.2018.2847625.

[112] K. M. Sunderland *et al.*, "The utility of multivariate outlier detection techniques for data quality evaluation in large studies : an application within the ONDRI project," *BMC Med Res Methodol*, vol. 5, pp. 1–16, 2019.

[113] N. Tasnim, M. K. Paul, and A. H. M. Sattar, "Identification of Drop Out Students Using Educational Data Mining," *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–5, 2019.

[114] H. Xu, Y. Wang, and Y. Wang, "Exploring a High-quality Outlying Feature Value Set for Noise-Resilient Outlier Detection in Categorical Data," *Session 1A: Content Understanding*, no. i, pp. 17–26, 2018.

[115] S. Salehjee, "Teaching science through stories: mounting scientific enquiry," *Early Child Dev Care*, vol. 190, pp. 1–12, 2020.

[116] M. Wook, N. A. M. Razali, S. Ramil, N. A. Wahab, and E. Al., "Opinion mining technique for developing student feedback analysis system using lexicon-based approach ( OMFeedback )," *Educ Inf Technol (Dordr)*, 2019.

[117] M. Misuraca, G. Scepi, and M. Spano, "Using Opinion Mining as an educational analytic: an integrated strategy for the analysis of students' feedback," *Studies in Educational Evaluation*, 2021.

[118] F. F. Balahadia, F. Ma. C. G, and I. C. Juanatas, "Teacher ' s Performance Evaluation Tool Using Opinion Mining with Sentiment Analysis," *IEEE Region 10 Symposium (TENSYMP)*, pp. 95–98, 2016.

[119] K. Z. Aung and N. M. Myo, "Sentiment Analysis of Students ' Comment Using Lexicon Based Approach," *IEEE computer society*, pp. 149–154, 2017.

[120] H. Khosravi, S. Buckingham, G. Chen, C. Conati, Y.-S. Tsai, and J. Kay, "Computers and Education : Artificial Intelligence Explainable Artificial Intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 3, no. May, 2022, doi: 10.1016/j.caeai.2022.100074.

[121] O. Ali, P. A. Murray, M. Momin, Y. K. Dwivedi, and T. Malik, "The effects of artificial intelligence applications in educational settings: Challenges and strategies," *Technol Forecast Soc Change*, vol. 199, Feb. 2024, doi: 10.1016/j.techfore.2023.123076.

[122] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, and Z. Du, "Artificial intelligence in education: A systematic literature review," Oct. 15, 2024, *Elsevier Ltd*. doi: 10.1016/j.eswa.2024.124167.

[123] A. hsuan S. Ma, "Faculty Members' Perspective on the Internationalization of Higher Education in Taiwan," *Asia Pacific Education Review*, 2024, doi: 10.1007/s12564-024-10016-z.

[124] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades," *Artif Intell Rev*, vol. 37, no. 4, pp. 331–344, Apr. 2012, doi: 10.1007/s10462-011-9234-x.

[125] F. Ouyang, M. Wu, L. Zheng, L. Zhang, and P. Jiao, "Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, Dec. 2023, doi: 10.1186/s41239-022-00372-4.

[126] X. Chen, H. Xie, and G. J. Hwang, “A multi-perspective study on Artificial Intelligence in Education: grants, conferences, journals, software tools, institutions, and researchers,” *Computers and Education: Artificial Intelligence*, vol. 1, no. August, p. 100005, 2020, doi: 10.1016/j.caeai.2020.100005.

[127] S. Tsai, C. Chen, Y. Shiao, J. Ciou, and T. Wu, “Precision education with statistical learning and deep learning : a case study in Taiwan,” *International Journal of Educational Technology in Higher Education*, 2020.

[128] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, and R. Nawaz, “Predicting Academic Performance of Students from VLE Big Data using Deep Learning Models,” *Comput Human Behav*, p. 106189, 2019.

[129] P. Kokol, M. Kokol, and S. Zagoranski, “Machine learning on small size samples: A synthetic knowledge synthesis,” *Sci Prog*, vol. 105, no. 1, pp. 1–16, 2022, doi: 10.1177/00368504211029777.

[130] M. D. Nemesure, M. V. Heinz, R. Huang, and N. C. Jacobson, “Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence,” *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-81368-4.

[131] S. Agnafors, M. Barmark, and G. Sydsjö, “Mental health and academic performance: a study on selection and causation effects from childhood to early adulthood,” *Soc Psychiatry Psychiatr Epidemiol*, vol. 56, no. 5, pp. 857–866, 2021, doi: 10.1007/s00127-020-01934-5.

[132] A. P. Delfino, “Student engagement and academic performance of students of Partido State University,” *Asian Journal of University Education*, vol. 15, no. 1, pp. 22–41, 2019, doi: 10.24191/ajue.v15i3.05.

[133] L. Cohausz, A. Tschalzev, Andrej. T.-M. De, and H. Stuckenschmidt, “Investigating the Importance of Demographic Features for EDM-Predictions,” in *Proceedings of the 16th International Conference on Educational Data Mining*, 2023. doi: 10.5281/zenodo.8115647.

[134] S. Blignaut *et al.*, “Internet Addiction and Academic Fatigue, Engagement, and Performance Among Undergraduates in Southwestern Universities in Nigeria,” *Higher Learning Research Communications*, vol. 2025, no. 1, pp. 1–23, 2025, doi: 10.18870/hlrc.v15i1.1553.

[135] C. Diamantini, L. Genga, A. Mircoli, D. Potena, and N. Zannone, “Understanding the stumbling blocks of Italian higher education system: A process mining approach,” *Expert Syst Appl*, vol. 242, May 2024, doi: 10.1016/j.eswa.2023.122747.

[136] O. Arifudin, U. Borneo Tarakan, S. Excelsius Surabaya, and U. Primagraha, “The Relationship Between Classroom Environment, Teacher Professional Development,

And Student Academic Performance In Secondary Education," *International Education Trend Issues*, vol. 2, no. 2, pp. 151–159, 2024, [Online]. Available: https://ijble.com/index.php/ieti

[137] C. C. Foong, P. Y. Liew, and A. J. Lye, "Changes in motivation and its relationship with academic performance among first-year chemical engineering students," *Education for Chemical Engineers*, vol. 38, no. August 2021, pp. 70–77, 2022, doi: 10.1016/j.ece.2021.11.002.

[138] S. M. Martinez, E. A. Frongillo, C. Leung, and L. Ritchie, "No food for thought: Food insecurity is related to poor mental health and lower academic performance among students in California's public university system," *J Health Psychol*, vol. 25, no. 12, pp. 1930–1939, 2020, doi: 10.1177/1359105318783028.

[139] G. S. Gowri, R. Thulasiram, and M. A. Baburao, "Educational Data Mining Application for Estimating Students Performance in Weka Environment," *IOP Conf Ser Mater Sci Eng*, pp. 1–10, 2017, doi: 10.1088/1757-899X/263/3/032002.

[140] P. Kamal and S. Ahuja, "Academic Performance Prediction Using Data Mining Techniques : Identification of Influential Factors Effecting the Academic Performance in Undergrad," *Advances in Intelligent Systems and Computing*, pp. 835–843, 2019, doi: 10.1007/978-981-13-0761-4.

[141] J. de D. Guterres, kusuma A. Laksitowening, and F. Sthevanie, "PROGRAMMING IN AN E-LEARNING ENVIRONMENT USING DECISION," *Syntax Literate : Jurnal Ilmiah Indonesia*, vol. 7, no. 1, 2022.

[142] F. Makombe and M. Lall, "A Predictive Model for the Determination of Academic Performance in Private Higher Education Institutions," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, pp. 415–419, 2020.

[143] S. Slater, S. Joksimovic, V. Kovanovic, and R. S. Baker, "Tools for Educational Data Mining : A Review," *Journal of Educational and Behavioral Statistics*, vol. 42, no. 1, pp. 85–106, 2017, doi: 10.3102/1076998616666808.

[144] J. Jacob, K. Jha, P. Kotak, and S. Puthran, "Educational Data Mining techniques and their applications," *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIoT 2015*, no. October 2015, pp. 1344–1348, 2015, doi: 10.1109/ICGCIoT.2015.7380675.

[145] C. Jalota and Dr. R. Agrawal, "Analysis of Educational Data Mining using classification," *IEEE, International Conference on Machine Learning, Big Data,* 2019.

[146] M. Al-Saleem, N. Al-Kathiry, S. Al-Osimi, and G. Badr, "Mining educational data to predict students' academic performance," *International publishing Switerland*, vol. 9166, pp. 403–414, 2015, doi: 10.1007/978-3-319-21024-7_28.

[147] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowl Based Syst*, vol. 200, Jul. 2020, doi: 10.1016/j.knosys.2020.105992.

[148] A. Acharya and D. Sinha, "Application of Feature Selection Methods in Educational Data Mining," *Int J Comput Appl*, vol. 103, no. 2, pp. 34–38, 2014, doi: 10.5120/18048-8951.

[149] R. S. Baker, "Challenges for the Future of Educational Data Mining : The Baker Learning Analytics Prizes," *Journal of Educational Data Mining*, vol. 11, no. 1, pp. 1–17, 2019.

[150] D. J. Lemay, C. Baek, and T. Doleck, "Comparison of learning analytics and educational data mining: A topic modeling approach," *Computers and Education: Artificial Intelligence*, vol. 2, no. March, p. 100016, 2021, doi: 10.1016/j.caeai.2021.100016.

[151] K. Mahawar and P. Rattan, "Optimizing Educational Outcome Prediction: Leveraging Chi-Square Method for Effective Feature Selection," in *IEEE 5th India Council International Subsections Conference (INDISCON)*, Springer Science and Business Media LLC, Feb. 2024, pp. 1–6. doi: 10.1007/s43681-020-00002-7.

[152] M. Ragab, A. M. K. Abdel Aal, A. O. Jifri, and N. F. Omran, "Enhancement of Predicting Students Performance Model Using Ensemble Approaches and Educational Data Mining Techniques," *Wirel Commun Mob Comput*, vol. 2021, 2021, doi: 10.1155/2021/6241676.

[153] M. Xiao, M. Zuo, X. Liu, K. Wang, and H. Luo, "After-School Behaviors, Self-Management, and Parental Involvement as Predictors of Academic Achievement in Adolescents," *Behavioral Sciences*, vol. 15, no. 2, Feb. 2025, doi: 10.3390/bs15020172.

[154] K. J. Bernard and P. E. Dzandza, "Effect of social media on academic performance of students in Ghanaian Universities: A case study of University of Ghana, Legon," *Library Philosophy and Practice*, vol. 2018, no. February, 2018.

[155] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Comput Educ*, p. 103676, 2019, doi: 10.1016/j.compedu.2019.103676.

[156] M. Saqr and A. Alamro, "The role of social network analysis as a learning analytics tool in online problem based learning," *BMC Med Educ*, vol. 19, no. 1, pp. 1–11, 2019, doi: 10.1186/s12909-019-1599-6.

[157] A. M. Hamid and B. M. Beli, "Influence of Social Networking Sites on the Academic Performance of Students of Select Higher Institutions of Learning in Bauchi, Nigeria," *GVU Journal of Communication Studies*, vol. 2, pp. 67–76, 2020.

[158] J. A. Middleton *et al.*, "A Social Network Analysis of Engineering Faculty Connections : Their Impact on Faculty Student-Centered Attitudes and Practices," *Educ Sci (Basel)*, 2022.

[159] N. M. Khan, M. Noreen, M. Hussain, and A. L. Hussaini, "The Impact of Cooperative Learning on Students' Academic Achievement and Social Behavior," *Harf-o-Sukhan*, vol. 8, no. 1, pp. 339–448, 2024.

[160] O. Kassak, M. Kompan, and M. Bielikova, "Student behavior in a web-based educational system: Exit intent prediction," *Eng Appl Artif Intell*, vol. 51, pp. 136–149, May 2016, doi: 10.1016/j.engappai.2016.01.018.

[161] L. Chen, L. Wang, and Y. Zhou, "Research on Data Mining Combination Model Analysis and Performance Prediction Based on Students' Behavior Characteristics," *Math Probl Eng*, vol. 2022, 2022, doi: 10.1155/2022/7403037.

[162] C. Brandmiller, K. Schnitzler, and H. Dumont, "Teacher perceptions of student motivation and engagement: longitudinal associations with student outcomes," *European Journal of Psychology of Education*, vol. 39, no. 2, pp. 1397–1420, Jun. 2024, doi: 10.1007/s10212-023-00741-1.

[163] G. Jiang, F. Liu, W. Liu, S. Liu, Y. Chen, and D. Xu, "Effects of information quality on information adoption on social media review platforms: moderating role of perceived risk," *Data Science and Management*, vol. 1, no. 1, pp. 13–22, Mar. 2021, doi: 10.1016/j.dsm.2021.02.004.

[164] A. A. A. Bukhari, H. H. Lean, W. A. A. Bukhari, S. Raza, and A. Pervaiz, "The dynamics of multidimensional educational poverty," *Data Science and Management*, Dec. 2024, doi: 10.1016/j.dsm.2024.03.004.

[165] A. M. Rabelo and L. E. Zárate, "A Model for Predicting Dropout of Higher Education Students," *Data Science and Management*, Jul. 2024, doi: 10.1016/j.dsm.2024.07.001.

[166] M. A. Camilleri, "Factors affecting performance expectancy and intentions to use ChatGPT: Using SmartPLS to advance an information technology acceptance framework," *Technol Forecast Soc Change*, vol. 201, Apr. 2024, doi: 10.1016/j.techfore.2024.123247.

[167] M. Kumar, K. Bajaj, B. Sharma, and S. Narang, "A Comparative Performance Assessment of Optimized Multilevel Ensemble Learning Model with Existing Classifier Models," *Big Data*, vol. 10, no. 5, pp. 371–387, 2021, doi: 10.1089/big.2021.0257.

[168] Sana, I. F. Siddiqui, and Q. A. Arain, "ANALYZING STUDENTS ' ACADEMIC PERFORMANCE THROUGH," *3C Tecnología. Glosas de innovación aplicadas a la pyme*, no. May, 2019.

[169] V. Matzavela and E. Alepis, "Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100035, 2021, doi: 10.1016/j.caeai.2021.100035.

[170] M. A. A. Mahdy, "The Impact of COVID-19 Pandemic on the Academic Performance of Veterinary Medical Students," *Front Vet Sci*, vol. 7, no. October, pp. 1–8, 2020, doi: 10.3389/fvets.2020.594261.

[171] E. S. Atlam, A. Ewis, M. M. A. El-Raouf, O. Ghoneim, and I. Gad, "A new approach in identifying the psychological impact of COVID-19 on university student's academic performance," *Alexandria Engineering Journal*, vol. 61, no. 7, pp. 5223–5233, 2022, doi: 10.1016/j.aej.2021.10.046.

[172] R. Asad, S. Altaf, S. Ahmad, H. Mahmoud, S. Huda, and S. Iqbal, "Machine Learning-Based Hybrid Ensemble Model Achieving Precision Education for Online Education Amid the Lockdown Period of COVID-19 Pandemic in Pakistan," *Sustainability*, vol. 15, no. 6, p. 5431, Mar. 2023, doi: 10.3390/su15065431.

[173] M. S. Patterson and P. Goodson, "Social network analysis for assessing college-aged adults' health: A systematic review," *Journal of American College Health*, vol. 67, no. 1, pp. 59–67, 2019, doi: 10.1080/07448481.2018.1462820.

[174] F. Fang, "A Study on the Application of Data Mining Techniques in the Management of Sustainable Education for Employment," *Data Sci J*, vol. 22, no. 1, 2023, doi: 10.5334/dsj-2023-023.

[175] Sana, I. F. Siddiqui, and Q. A. Arain, "ANALYZING STUDENTS ' ACADEMIC PERFORMANCE THROUGH," *3C Tecnología. Glosas de innovación aplicadas a la pyme*, no. May, 2019.

[176] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, Springer Verlag, 2003, pp. 267–274. doi: 10.1007/978-3-540-45226-3_37.

[177] K. M. . Iraqi, "Predicting Students' Academic Performance Through Supervised Machine Learning," in *ICISCT'20 : 2nd International Conference on Information Science and Communication Technology*, IEEE, 2020.

[178] F. Al-Obeidat, A. Tubaishat, A. Dillon, and B. Shah, "Analyzing students' performance using multi-criteria classification," *Cluster Comput*, vol. 21, no. 1, pp. 623–632, May 2017, doi: 10.1007/s10586-017-0967-4.

[179] Maozhen. Li, "Application of Artificial Neural Networks for Prediction of Learning Performances," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery : ICNC-FSKD 2016*, IEEE, 2016, pp. 745–751.

[180] H. Al-Shehri, A. Al-Qarni, and L. Al-Saati, "Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor," in *2017 IEEE 30th CanadiConference on Electrical and Computer Engineering (CCECE)*, IEEE, 2017.

[181] Y. H. Hu, C. L. Lo, and S. P. Shih, "Developing early warning systems to predict students' online learning performance," *Comput Human Behav*, vol. 36, pp. 469–478, 2014, doi: 10.1016/j.chb.2014.04.002.

[182] C. Romero, M. I. López, J. M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Comput Educ*, vol. 68, pp. 458–472, 2013, doi: 10.1016/j.compedu.2013.06.009.

[183] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, and M. Hernandez, "Perspectives to Predict Dropout in University Students with Machine Learning," in *2018 IEEE International Work Conference on Bioinspired Intelligence, IWOBI 2018 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Sep. 2018. doi: 10.1109/IWOBI.2018.8464191.

[184] D. Alboaneen, M. Almelihi, R. Alsubaie, R. Alghamdi, L. Alshehri, and R. Alharthi, "Development of a Web-Based Prediction System for Students' Academic Performance," *Data (Basel)*, vol. 7, no. 2, Feb. 2022, doi: 10.3390/data7020021.

[185] F. Hajjej, S. Ayouni, M. A. Alohali, and M. Maddeh, "Novel Framework for Autism Spectrum Disorder Identification and Tailored Education With Effective Data Mining and Ensemble Learning Techniques," *IEEE Access*, vol. 12, pp. 35448–35461, 2024, doi: 10.1109/ACCESS.2024.3349988.

[186] H. Xu and M. Kim, "Combination prediction method of students' performance based on ant colony algorithm," *PLoS One*, vol. 19, no. 3 March, Mar. 2024, doi: 10.1371/journal.pone.0300010.

[187] M. Arifin, W. Widowati, and F. Farikhin, "Optimization of Hyperparameters in Machine Learning for Enhancing Predictions of Student Academic Performance," *Ingenierie des Systemes d'Information*, vol. 28, no. 3, pp. 575–582, Jun. 2023, doi: 10.18280/isi.280305.

[188] S. Chen and Y. Ding, "A Machine Learning Approach to Predicting Academic Performance in Pennsylvania's Schools," *Soc Sci*, vol. 12, no. 3, Mar. 2023, doi: 10.3390/socsci12030118.

[189] S. Albahli, "Efficient hyperparameter tuning for predicting student performance with Bayesian optimization," *Multimed Tools Appl*, vol. 83, no. 17, pp. 52711–52735, May 2024, doi: 10.1007/s11042-023-17525-w.

[190] R. Costa-Mendes, T. Oliveira, M. Castelli, and F. Cruz-Jesus, "A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach,"

*Educ Inf Technol (Dordr)*, vol. 26, no. 2, pp. 1527–1547, Mar. 2021, doi: 10.1007/s10639-020-10316-y.

[191] H. M. Nagy, W. M. Aly, and O. F. Hegazy, “An Educational Data Mining System for Advising Higher Education Students,” *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 7, no. 10, pp. 1266–1270, 2013.

[192] M. Bellaj, A. Ben Dahmane, and L. Sefian, “Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students’ Academic Performance,” *International journal of online and biomedical engineering*, vol. 20, no. 3, pp. 55–74, 2024, doi: 10.3991/ijoe.v20i03.46287.

[193] A. Z. Ye *et al.*, “High-Dimensional Feature Selection Based on Improved Binary Ant Colony Optimization Combined with Hybrid Rice Optimization Algorithm,” *International Journal of Intelligent Systems*, vol. 2023, 2023, doi: 10.1155/2023/1444938.

[194] B. Minaei-Bidgoli, D. A. Kashy ’, G. Kortemeyer’, and W. F. Punch, “PREDICTING STUDENT PERFORMANCE: AN APPLICATION OF DATA MINING METHODS WITH AN EDUCATIONAL WEB-BASED SYSTEM,” in *33rd IEEE Frontiers in Education Conference- Session T2A-13*, 2003.

[195] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student’ performance prediction using machine learning techniques,” *Educ Sci (Basel)*, vol. 11, no. 9, Sep. 2021, doi: 10.3390/educsci11090552.

[196] A. Algahtani, “A COMPARATIVE STUDY OF AI-BASED EDUCATIONAL TOOLS: EVALUATING USER INTERFACE EXPERIENCE AND EDUCATIONAL IMPACT,” *J Theor Appl Inf Technol*, vol. 15, no. 5, 2024, [Online]. Available: www.jatit.org

[197] A. Rashad Sayed, M. Helmy Khafagy, M. Ali, and M. Hussien Mohamed, “Predict student learning styles and suitable assessment methods using click stream,” *Egyptian Informatics Journal*, vol. 26, Jun. 2024, doi: 10.1016/j.eij.2024.100469.

[198] N. Demirci and B. Üniversitesi, “University Students’ Perceptions of Web-based vs. Paper-based Homework in a General Physics Course,” *Eurasia Journal of Mathematics*, vol. 3, no. 1, pp. 29–34, 2007.

[199] P. Sokkhey, P. Penh, and C. T. Okazaki, “Developing Web-based Support Systems for Predicting Poor-performing Students using Educational Data Mining Techniques,” *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 11, no. 7, 2020, [Online]. Available: www.ijacsa.thesai.org

[200] A. Cohen, “Analysis of student activity in web-supported courses as a tool for predicting dropout,” *Educational Technology Research and Development*, vol. 65, no. 5, pp. 1285–1304, Oct. 2017, doi: 10.1007/s11423-017-9524-3.

# LIST OF PUBLICATIONS

## JOURNALS

1. Mahawar K., Rattan P., "Dimensionality Reduction using Feature Selection Techniques on EDM for Student Academic Performance Prediction", International Journal on Recent and Innovation Trends in Computing and Communication, Nov 2023
2. Mahawar K., Rattan P., "Empowering education: Harnessing ensemble machine learning approach and ACO-DT classifier for early student academic performance prediction", Education and Information Technologies, Sept 2024
3. Mahawar K., Rattan P., "Employing Artificial Bee and Ant Colony Optimization in Machine Learning Techniques as a Cognitive Neuroscience Tool", Scientific Reports, March 2025

## CONFERENCES

1. Mahawar K., Rattan P., "A Comprehensive Study on Student's Academic Performance Prediction Using Artificial Intelligence", Recent Advances in Computing Sciences (RACS), 2023
2. Mahawar K., Rattan P., "A Comprehensive Investigation Into The Dimensions Of Educational Data Mining Using Artificial Intelligence", International Conference on Advanced Computing & Communication Technologies (ICACCTech), 2023
3. Mahawar K., Rattan P., "Early Prediction of Undergraduate Students Academic Achievement Through Machine Learning", $2^{nd}$ International Conference on Networks, Intelligence & Computing (ICONIC 2024), 2024
4. Mahawar K., Rattan P., "Optimizing Educational Outcome Prediction: Leveraging Chi-Square Method for Effective Feature Selection", 2024 IEEE $5^{th}$ India Council Subsections Conference (INDISCON), 2024

# Appendix I

# Journal

Mahawar K., Rattan P., "Dimensionality Reduction using Feature Selection Techniques on EDM for Student Academic Performance Prediction", International Journal on Recent and Innovation Trends in Computing and Communication, Nov 2023

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 10*
*Article Received: 26 August 2023 Revised: 20 October 2023 Accepted: 02 November 2023*

## Dimensionality Reduction using Feature Selection Techniques on EDM for Student Academic Performance Prediction

**Kajal Mahawar[1], Punam Rattan[2]**
[1]School of Computer Applications, Lovely Professional University, Phagwara (Punjab), India
kajalmahawar01@gmail.com
[2]School of Computer Applications, Lovely Professional University, Phagwara (Punjab), India
punamrattan@gmail.com

**Abstract—** The aim of this study is to perform dimensionality reduction on student datasets using different feature selection strategies in order to better forecast students' academic progress and minimize high-dimensionality concerns. The effectiveness of models built utilizing classification algorithms and pertinent features chosen using different feature selection strategies has been experimentally evaluated in this study. Nine feature selection methods, including Chi-Square, Pearson Correlation, Variance Threshold, Feature Importance, Recursive Feature Elimination (RFE), Lasso, Ridge, Random Forest, and XGBoost, as well as one classification method, logistic regression, were applied to the student dataset to determine the results of the analysis. In this experiment, dimensionality reduction techniques were used to reduce the number of features from 70 to the optimal 21 feature subsets for analysis of student academic performance prediction. Five measures were used to assess the effectiveness of the techniques: $r^2$, precision, recall, f-score, and accuracy. With the logistic regression classifier, the feature subset chosen by the chi-square had the highest accuracy of 72.4%, precision of 71.4%, recall of 72.4%, f-score of 70%, and r-square of -0.221%.

**Keywords**- Dimensionality reduction, Students' academic performance, Feature selection, Logistic regression, Measures

### I. INTRODUCTION

Student academic performance prediction is a study area having applications of deep learning, machine learning, artificial intelligence, and statistics for the education sector. Machine learning and data mining approaches employ to shift through data from educational settings in search of patterns and predictions that describe students' actions and performance. It enhances students' brains and personalities. A crucial component of the future perspective in education is the ability to predict academic accomplishment in advance. Since students are the main participants in educational institutions, these institutions can adapt to the shifting needs of the society by examining student data and drawing various conclusions from data. Additionally, the outcomes of projections might be useful for developing plans to raise educational standards. Instructors, students, and surroundings are important sorts of participants in an academic institution. These participants' interactions provide a large amount of data that can be methodically grouped to harvest priceless knowledge. Predicting a student's success is a major concern for school administrators and the government. It could provide suitable warnings to students who are in danger by anticipating their grades and assisting them in avoiding problems and overcoming any study challenges. As a result, a range of machine learning-related algorithms have been developed [1]. The user can handle and evaluate the student's data acquired from many sources thanks to a developing study field called "mining of student's data with machine learning." An author used a meta-analysis to assess the overall efficacy of interactive white board-based training by finding empirical studies that appear on students' cognitive learning results [2]. Another author [3] experimented questioned college students in China who used blended learning for the first time in the semester of 2020 and 2021 to discover the significant features that influence beginners' intention. The associations between important technological variables and connected classroom climate in cloud classrooms are empirically supported according to the authors [4]. According to the [5], student academic performance prediction from their previous performance record is a challenging problem in the education sector because of student's microarray data.

There are certain distinctive features in the microarray data. In other words, there are few samples and a high degree of dimension. Dimensionality reduction is a strategy for transforming a dataset with multiple dimensions into a dataset with fewer dimensions while maintaining the same amount of information. It is a vital preprocessing step in the classification

# Appendix II

# Journal

Mahawar K., Rattan P., “Empowering education: Harnessing ensemble machine learning approach and ACO-DT classifier for early student academic performance prediction”, Education and Information Technologies, Sept 2024

Education and Information Technologies
https://doi.org/10.1007/s10639-024-12976-6

## Empowering education: Harnessing ensemble machine learning approach and ACO-DT classifier for early student academic performance prediction

Kajal Mahawar[1] · Punam Rattan[1]

Received: 19 February 2024 / Accepted: 9 August 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

**Abstract**
Higher education institutions have consistently strived to provide students with top-notch education. To achieve better outcomes, machine learning (ML) algorithms greatly simplify the prediction process. ML can be utilized by academicians to obtain insight into student data and mine data for forecasting the performance. In this paper, the authors proposed an ML-based student prediction model based on the demographic, social, psychological, and economic factors, collectively. The dataset utilized for this study was compiled from a designed questionnaire administered to second-year undergraduate students. The objective of this study is to uncover factors that could assist in predicting students' performance. Eight ML classifiers, logistic regression, random forest, support vector machine, XGBoost, support vector machine with a linear kernel, naïve Bayes, K-Nearest Neighbor, and decision tree are used to forecast student performance. Additionally, nine feature selection techniques, variance threshold, XGBoost, feature importance, recursive feature elimination, chi-square, ridge, Pearson correlation, lasso, and random forest, are employed to determine optimal factors. The authors experimented with each technique by creating two sets of training and testing data with 80:20 and 70:30 proportions, respectively. Comparatively, the ensemble DXK (DT + XGB + KNN) model with cross-validation and 80:20 proportions outperformed other standard classifiers, achieving a highest accuracy of 97.83%, an r-square of 96.17%, a precision of 97.94%, a recall of 97.83%, and an f1-score of 97.88%. These were the highest among all models tested. Additionally, the authors propose the ACO-DT model, which improves the prediction performance of the top-performing DT classifier by utilizing the Ant Colony Optimization technique. The findings demonstrate that the proposed model with 80:20 proportions achieve an accuracy of 98.15%, an f1-score of 98.16%, a precision of 98.18%, a recall of 98.15%, and an r-square of 84.75%, surpassing all other models for forecasting student performance. Using the specified data size, this model creation time is 8.49 s. The authors also recommended the future research directions to further enhance this study.

Extended author information available on the last page of the article

Published online: 02 September 2024 Springer

# Appendix III

# Journal

Mahawar K., Rattan P., "Employing Artificial Bee and Ant Colony Optimization in Machine Learning Techniques as a Cognitive Neuroscience Tool", Scientific Reports, March 2025

**scientific** reports

Check for updates

OPEN

## Employing artificial bee and ant colony optimization in machine learning techniques as a cognitive neuroscience tool

Kajal Mahawar[1], Punam Rattan[1], Ammar Jalamneh[2], Mohd Shukri Ab Yajid[3], Omar Abdeljaber[4], Raman Kumar[5,6], Ayodele Lasisi[7] & Muhammad Imam Ammarullah[8,9]

**Higher education is essential because it exposes students to a variety of areas. The academic performance of IT students is crucial and might fail if it isn't documented to identify the features influencing them, as well as their strengths and shortcomings. The student academic prediction system needs to be enhanced so that teachers can forecast their students' performance. Numerous studies have been conducted to increase the prediction accuracy of IT students, but they encountered difficulties with unbalanced data and algorithm tuning. To address these issues, the study proposed different machine learning (ML) algorithms that handled imbalanced data by applying the synthetic minority oversampling technique (SMOTE) and employing hyperparameter tuning algorithms to enhance prediction during the training process. The ML models we used were decision tree (DT), k-nearest neighbor, and XGBoost. The models were fine-tuned by applying Ant colony optimization (ACO) and artificial bee colony optimization techniques. Subsequently, these optimization techniques further enhanced the performance of the models. After comparing them, the results showed that SMOTE and ACO combined with the DT model outperformed other models for academic prediction. Additionally, the study utilized the Kendall Tau correlation coefficient technique to analyze the correlation between features and identify factors that positively or negatively impact student success.**

**Keywords** Student learning outcomes, Machine learning, Hyperparameter tuning, Decision tree, Artificial bee colony, Ant colony optimization, Kendall Tau

Indeed, one of the most challenging and extensively researched areas in machine learning (ML) revolves around modeling student performance[1]. Predicting IT students' academic achievements is pivotal for educational planning and decision-making. Tailoring ML techniques to address the distinct challenges faced by these students offers promising avenues to enhance predictive accuracy and optimize educational outcomes. According to higher education studies, the high attrition rate demonstrates the ineffectiveness of the prior educational initiatives. Significant reforms in higher education are required to address the problem, increase student retention, and raise graduation rates. The crucial stage at which the research concentrated on the features significantly affecting the outcomes was performance prediction. In addition, the prediction models inside the designated domain experienced low efficacy and precision, necessitating modifications to yield superior outcomes suitable for real-time analysis. Nonetheless, for the decision-maker to effectively manage their student, academic prediction from the student needs to be more accurate[2].

[1]Lovely Professional University, Phagwara 144411, Punjab, India. [2]College of Arts and Science, Applied Science University, Al Ekir 3201, Kingdom of Bahrain. [3]Management and Science University, Shah Alam 40100, Selangor, Malaysia. [4]Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan. [5]Department of Mechanical and Production Engineering, Guru Nanak Dev Engineering College, Ludhiana 141006, Punjab, India. [6]Jadara University Research Center, Jadara University, Irbid 733, Jordan. [7]Department of Computer Science, College of Computer Science, King Khalid University, Abha 61421, Asir, Saudi Arabia. [8]Department of Mechanical Engineering, Faculty of Engineering, Universitas Diponegoro, Semarang 50275, Central Java, Indonesia. [9]Undip Biomechanics Engineering & Research Centre (UBM-ERC), Universitas Diponegoro, 50275 Semarang, Central Java, Indonesia. email: imamammarullah@gmail.com

# Appendix IV

# Conference

Mahawar K., Rattan P., "A Comprehensive Study on Student's Academic Performance Prediction Using Artificial Intelligence", Recent Advances in Computing Sciences (RACS), 2023

LOVELY PROFESSIONAL UNIVERSITY

*Transforming Education Transforming India*

RACS

SCHOOL OF COMPUTER APPLICATION

Certificate No. 256843

## Certificate of Paper Presentation

This is to certify that **Dr./Mr./Ms. Kajal Mahawar** of **Lovely Professional University, Punjab** has presented a paper entitled "**A Comprehensive Study on Student's Academic Performance Prediction Using Artificial Intelligence**" in the 1st **International Conference on Recent Advances in Computing Sciences (RACS-2022)** held from 4th to 5th November 2022, organized by School of Computer Application at Lovely Professional University, Punjab.

Date of Issue : 30-11-2022
Place : Phagwara (Punjab), India

| Prepared by (Administrative Officer-Records) | Convener (RACS-2022) | Conference Chair (RACS-2022) | Session Chair (RACS-2022) |
|---|---|---|---|

# Appendix V

## Conference

Mahawar K., Rattan P., “A Comprehensive Investigation Into The Dimensions Of Educational Data Mining Using Artificial Intelligence”, International Conference on Advanced Computing & Communication Technologies (ICACCTech), 2023


#123

SVIET

SWAMI VIVEKANAND GROUP OF INSTITUTES

ICACCTech 2023

INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING & COMMUNICATION TECHNOLOGIES

Certificate

This certificate is proudly presented to Dr./ Mr./ Ms. Kajal Mahawar, Lovely Professional University, India, for the presentation of the paper titled "A Comprehensive Investigation Into The Dimensions Of Educational Data Mining Using Artificial Intelligence" at the International Conference on Advanced Computing & Communication Technologies (ICACCTech 2023) organized jointly by Swami Vivekanand Institute of Engineering and Technology & B.M. Institute of Engineering & Technology, INDIA on 23th-24th December 2023.

Dr. Naveen Garg
Conference General Chair

IEEE COMPUTER SOCIETY
CONFERENCE PUBLISHING SERVICES

Prof. (Dr) Harish Mittal
Conference Program Chair

# Appendix VI

# Conference

Mahawar K., Rattan P., "Early Prediction of Undergraduate Students Academic Achievement Through Machine Learning", 2$^{nd}$ International Conference on Networks, Intelligence & Computing (ICONIC 2024), 2024

ICONIC
International Conference on Networks, Intelligence and Computing

LPU SCHOOL OF COMPUTER APPLICATION

NAAC GRADE A++

Certificate No. 327566

## Certificate of Presentation

This is to certify that **Dr./Mr./Ms. Kajal Mahawar** of **Lovely Professional University** has given Oral/Poster Presentation on **Early Prediction of Undergraduate Students Academic Achievement Through Machine Learning** in the **2nd International Conference on Networks, Intelligence & Computing (ICONIC 2024)** held from **26th** to **27th April 2024**, organized by School of Computer Application, Lovely Faculty of Technology and Sciences at Lovely Professional University, Punjab.

Date of Issue : 29-05-2024
Place : Phagwara (Punjab), India

Prepared by
(Administrative Officer-Records)

Convener
(ICONIC-2024)

Head of Faculty
Lovely Professional University

# Appendix VII

# Conference

Mahawar K., Rattan P., "Optimizing Educational Outcome Prediction: Leveraging Chi-Square Method for Effective Feature Selection", 2024 IEEE 5$^{th}$ India Council Subsections Conference (INDISCON), 2024

## Optimizing Educational Outcome Prediction: Leveraging Chi-Square Method for Effective Feature Selection


Kajal Mahawar
*School of Computer Applications*
*Lovely Professional University*
Phagwara, India
kajalmahawar01@gmail.com

Punam Rattan
*School of Computer Applications*
*Lovely Professional University*
Phagwara, India
punamrattan@gmail.com



2024 IEEE 5th India Council International Subsections Conference (INDISCON) | 979-8-3503-7675-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/INDISCON62179.2024.10744404



*Abstract*—**This study investigates the use of machine learning (ML) methods to predict the academic performance of second-year undergraduate students based on data collected from a 2021-2022 questionnaire. The Chi-Square feature selection method was utilized alongside eight ML classifiers: Support Vector Machine with Linear Kernel, Random Forest, Logistic Regression, Naive Bayes, K-Nearest Neighbor, Decision Tree, XGBoost, and another Support Vector Machine. Each classifier was tested with training and testing data sets split in 60:40 and 90:10 ratios. Performance metrics such as $r^2$, precision, recall, f-score, and accuracy were used to assess their effectiveness. The Decision Tree classifier, tested with a 90:10 data split, achieved the highest accuracy at 91.70%, outperforming the other methods. This demonstrates the potential of ML techniques, particularly decision trees, to accurately predict student academic performance, thus contributing to improved educational practices in higher education.**



*Keywords- Machine learning, Student academic performance, Chi-Square, ML classifiers, Decision tree, Educational practices*


### I. INTRODUCTION

Student academic success is crucial for a nation's economic and social progress. Successful students often become influential leaders and valuable contributors to society, driving social advancement. Assessing student achievement has been a major focus of past research, recognized as a complex area influenced by various factors. Despite these factors varying among students and locations, they significantly affect academic outcomes [1]. The effectiveness of teaching is gauged by how well it meets students' needs, making it crucial for higher education institutions to anticipate students' performance [2]. Alternatively, Machine learning is extensively utilized for predictions in areas such as astronomy, commerce, and education. Educational institutions are particularly interested in creating models to predict student performance, contributing significantly to the growing body of research in artificial intelligence [3]. According to [4], the convergence of student academic performance and machine learning classifiers has attracted growing interest in educational research. With the rise of big data and technological progress, educators and researchers are investigating new methods to better understand the various factors influencing student success [5]. Utilizing extensive datasets that include demographic, socio-economic, behavioral, and learning style information, machine learning classifiers can transform our comprehension of academic success. These tools can guide specific interventions to enhance student learning and development [6]. Recently, the use of artificial intelligence in educational data mining has garnered significant interest from researchers. As technology progresses, both students and teachers utilize various AI-driven tools and applications, including training robots, intelligent tutoring systems, and adaptive learning platforms [7].

Predicting student academic performance involves using data-driven methods to estimate the likelihood of students succeeding in their studies [8]. By examining a wide range of factors, including demographics, socioeconomic status, past academic performance, attendance, behavior, health, relationships, time management, context, and available educational resources, predictive models can be crafted to anticipate student outcomes. The value of predicting student academic performance is in its ability to guide targeted interventions and support strategies to enhance student success. Early identification of at-risk students allows educators to take proactive steps to address challenges and offer needed support. Subsequently, according to [9], predictive models allow educational institutions to distribute resources more effectively, improve educational programs, and boost overall student success grades.

The paper is organized as follows: Section 2 covers the literature review, and Sections 3 and 4 describe the materials, methods, and experiments. Section 5 presents the results and discussion. Lastly, Section 6 provides the study's conclusions and suggests directions for future research.

### II. LITERATURE REVIEW

Improving the outcomes of traditional machine learning algorithms involves employing various techniques tailored to technology's effectiveness. For instance, ensemble learning stands out as a highly effective paradigm for data classification. Despite the recent popularity and widespread use of deep learning, ensemble learning algorithms are often preferred over deep learning for solving various supervised learning tasks due to the extensive time and data required to train and test deep learning models [3]. While it is known that factors like interpersonal skills, test anxiety, and academic stress affect student performance, gender differences have also been observed in relation to these factors. Further exploration of these elements across genders is warranted [4].

The researchers studied the optimal ML classifier for developing intelligent systems that can predict undergraduate research areas based on academic backgrounds and job market trends [5]. Moreover, the study referenced in [6] aimed to assess the effectiveness of ML models in predicting student academic performance within specific research domains. Likewise, the primary goal of the study [8] was to enhance and implement an hybrid model to fetch learners at-risk of academic challenges in the epidemic.

Employing a logistic regression classifier, [9] researchers predicted the academic performance of postgraduate students based on their feedback. They employed another machine learning model to analyze how students' learning behaviours influenced their academic outcomes. In [10], researchers utilized Naive Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and ID3 machine learning techniques. They improved classifier performance through ensemble methods. For predicting student dropout from university courses [11], researchers proposed an algorithm integrating Random Forest (RF), Gradient Boosting (GB), and XGBoost. The aim of the study mentioned in [12] was to


Authorized licensed use limited to: Vivekanand Education Society's Inst of Tech. Downloaded on December 05,2024 at 07:22:20 UTC from IEEE Xplore. Restrictions apply.

# Appendix VIII

## List of the Institutions

| | |
|---|---|
| **1.** | St. Aloysius' College (Autonomous), Jabalpur (M.P.) |
| **2.** | Government Arts College, Panagar (M.P.) |
| **3.** | Rajeev Gandhi Institute of Science and Technology, Jabalpur (M.P.) |
| **4.** | Navyug Arts & Commerce College, Jabalpur (M.P.) |
| **5.** | Ayodhya Prasad Narmada College, Jabalpur (M.P.) |

# Appendix IX
# A Copy of Original Questionnaire

# A SURVEY ON ACADEMIC PERFORMANCE OF THE STUDENTS

Dear Student's,
This is part of my Ph.D. work. I humbly request you to fill-up this questionnaire. Your contribution in this regard will be highly appreciated and the given information will be used for academic purpose only and will be kept confidential. Hope filling this questionnaire will bring insight for both researcher and respondents. It is divided into five parts.

**PART - I DEMOGRAPHIC PROFILE**
Note: Please make a tick in appropriate box provided against each statement.

* Indicates required question

1. Email *

2. Name of the respondent *

3. Course Name *

*Mark only one oval.*

- BCA
- B.com CA
- B.Sc CS
- Any Other

4. Institution Name *

5. Gender of the Respondent *

*Mark only one oval.*

- Male
- Female
- Transgender

6. Age of the Respondent *

*Mark only one oval.*

- 18 years
- 19-25 years
- Above 25 years

7. Respondent category *

*Mark only one oval.*

- Unreserved (Gen)
- EWS
- OBC- Creamy
- OBC- Non Creamy
- SC
- ST

8. Respondent Religion *

*Mark only one oval.*

- Hindu
- Muslim
- Christian
- Jain
- Sikh
- Other

9. Marital Status *

*Mark only one oval.*

- Married
- Unmarried
- Divorced
- Widow/Widower

10. The number of children's *

*Mark only one oval.*

- 1 to 2
- 3 to 5
- More than 5
- None

11. Respondent Other Information *

*Mark only one oval per row.*

| | Yes | No |
|---|---|---|
| **Are you a person with disability?** | | |
| **Are you the first child?** | | |
| **Are you doing any part-time job?** | | |
| **Are you a single girl child?** | | |

12. Area you live? *

*Mark only one oval.*

- Urban
- Rural

13. Are you a single parent child? *

*Mark only one oval.*

- Yes *Skip to question 14*
- No *Skip to question 14*

**PART - II FAMILY INFORMATION**

14. Family Academic Qualification *

*Mark only one oval per row.*

| | Below Matric Pass | Matric Pass (10th pass) | 12th pass | Graduate | Post Graduate | Higher Qualification than Post Graduate |
|---|---|---|---|---|---|---|
| **Father/Husband Academic Qualification** | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ |
| **Mother/Wife Academic Qualification** | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ |

15. Family Occupation *

*Mark only one oval per row.*

| | Government employee | Private sector employee | Business | Others (farmer, venders etc.) | Retired | None |
|---|---|---|---|---|---|---|
| **Father/ Husband occupation** | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ |
| **Mother/ Wife occupation** | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ |

16. Total Income (Father or Husband and/or Mother or Wife) *

*Mark only one oval.*

- ⬭ Below 50,000
- ⬭ 50,000 to 1 lakh
- ⬭ 1 lakh to 2 lakh
- ⬭ 2 lakh to 5 lakh
- ⬭ Above 5 lakh

17. On a scale of 1 to 5, how much do you think your family environment can help you in your academics from following concerns? (1 being the lowest and 5 being the highest) *

*Mark only one oval per row.*

| | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| **Help, trust and financial stability of family** | ◯ | ◯ | ◯ | ◯ | ◯ |
| **Support for open expression of emotions** | ◯ | ◯ | ◯ | ◯ | ◯ |
| **Family conflicts** | ◯ | ◯ | ◯ | ◯ | ◯ |
| **Freedom** | ◯ | ◯ | ◯ | ◯ | ◯ |
| **Moral values and motivation** | ◯ | ◯ | ◯ | ◯ | ◯ |

18. Family Type *

*Mark only one oval.*

◯ Nuclear

◯ Joint

## PART - III ACADEMIC INFORMATION

19. Respondent Course type *

*Mark only one oval.*

◯ Practical

◯ Non-Practical

20. Respondent admission type *

*Mark only one oval.*

- Regular
- Private

21. Other academic information *

*Mark only one oval per row.*

| | High | Medium | Low |
|---|---|---|---|
| **Are you involved in other co-curricular activities in college?** | | | |
| **Respondent engagement in class** | | | |

22. Respondent 1st year percentage (No *round off)* *

*Mark only one oval.*

- more than 60 %
- less than 60%

23. Attendance *

*Mark only one oval.*

- High (more than 75 % )
- Medium (55 % to 74 %)
- Low ( less than 55 %)

24. Distance from home to college *

*Mark only one oval.*

- less than 15 minutes
- 15 to 30 minutes
- 30 minutes to 1 hour
- 1 hour to 1.30 hours
- more than 1.30 hours

25. Other Non-Academic Information *

*Mark only one oval per row.*

| | Yes | No |
|---|---|---|
| **Did you face any kind of difficulty to get the transportation services from your home to college or vice versa?** | | |
| **Did your institution provide any kind of transportation facility for students?** | | |
| **Are you happy with your current course/program choice?** | | |
| **Are you getting any kind of scholarship from your institution?** | | |

26. Are you a Hostler? *

*Mark only one oval.*

- Yes
- No

27. Respondent Learning mode *

*Mark only one oval per row.*

| | Online mode | Offline mode |
|---|---|---|
| **Respondent Learning mode preference** | | |
| **Respondent current Learning mode** | | |

28. Will your course choice help you grow in your professional interest? *

*Mark only one oval.*

- Yes
- No

29. Respondent relationship with people at institution *

*Mark only one oval per row.*

| | Excellent | Very Good | Good | Fair | Bad |
|---|---|---|---|---|---|
| **Relationship with classmates** | | | | | |
| **Relationship with teachers/mentors** | | | | | |
| **Relationship with administration staff** | | | | | |
| **How would you evaluate the quality of academic facility you received at your pass out school?** | | | | | |
| **How would you evaluate the quality of academic facility you received at your institution?** | | | | | |

## PART - IV OTHER INFORMATION

30. Did you face any kind of trauma in your childhood? *

*Mark only one oval.*

- Yes
- No

31. Number of hours spends in coaching classes? *

*Mark only one oval.*

- 1 hour
- 1 to 2 hours
- 2 to 4 hours
- 4 hours
- None

32. Number of siblings *

*Mark only one oval.*

- none
- 1-2
- 3-5
- more than 5

33. Respondent other activities *

*Mark only one oval per row.*

| | Yes | No |
|---|---|---|
| **Are you aware about importance of education in once life?** | ⬭ | ⬭ |
| **Are you aware about importance of career in once life?** | ⬭ | ⬭ |
| **Are you facing any gender inequality in your surroundings (family, college, or other places)?** | ⬭ | ⬭ |
| **Have your past activities negatively affected your academic performance?** | ⬭ | ⬭ |
| **Are you involved in other co-curricular activities in home?** | ⬭ | ⬭ |
| **Do you have any bad habits such as alcohol consumption, smoking etc.?** | ⬭ | ⬭ |
| **Did you suffer any health** | ⬭ | ⬭ |

| | | |
|---|---|---|
| issues? | | |
| Did your surroundings motivated and encouraged you towards your studies? | ⬭ | ⬭ |
| Did covid-19 pandemic affect your learning ability? | ⬭ | ⬭ |
| Are you involved in any illegal activity that will affect your academic performance? | ⬭ | ⬭ |
| Are you struggling with personal problems that are negatively affecting your academic performance? | ⬭ | ⬭ |

34. Do you face any kind of language barrier in your studies? *

*Mark only one oval.*

- Yes
- No

35. Active hour rate spending in *

*Mark only one oval per row.*

| | None | less than 1 hour | 1 to 2 hours | 2 to 4 hours | more than 4 hours |
|---|---|---|---|---|---|
| **social sites platforms ( instagram, facebook, web series, whatsapp etc.) for fun** | | | | | |
| **online video games (freefire, pubG etc.)** | | | | | |
| **other places such as canteen, café, campus etc.** | | | | | |
| **study places such as classrooms, lab, library etc.** | | | | | |
| **self study** | | | | | |

36. How much number do you give yourself out of 5 (highest 5, lowest 1) in: *

*Mark only one oval per row.*

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Motivated towards studies** | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ |
| **Confident towards pupils** | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ |
| **Emotional stability** | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ |

## PART - V FRIENDS INFORMATION

37. Friend other information *

*Mark only one oval per row.*

| | High | Medium | Low |
|---|---|---|---|
| **Your friend engagement in class** | ⬭ | ⬭ | ⬭ |
| **Your friend's attendance** | ⬭ | ⬭ | ⬭ |

38. Friend's background *

*Mark only one oval per row.*

| | Yes | No |
|---|---|---|
| **Did your friend motivated, help or encouraged you to do studies?** | | |
| **Did your friend have any bad habits such as alcohol consumption, smoking etc.?** | | |

This content is neither created nor endorsed by Google.

Google Forms

# Appendix X
# A Copy of Revised Questionnaire

# A SURVEY ON ACADEMIC PERFORMANCE OF THE STUDENTS: REFRAMED Questionnaire

Dear Student's,
This is part of my Ph.D. work. I humbly request you to fill-up this questionnaire. Your contribution in this regard will be highly appreciated and the given information will be used for academic purpose only and will be kept confidential. Hope filling this questionnaire will bring insight for both researcher and respondents.
**Note: Please make a tick in appropriate box provided against each statement.**

* Indicates required question

1. Name of the respondent *

_______________________

2. Course Name *

*Mark only one oval.*

- BCA
- B.com CA
- B.Sc CS
- Any Other

3. Institution Name *

_______________________

4. 1. Gender of the Respondent *

*Mark only one oval.*

- Male
- Female
- Transgender

5. 2. Age of the Respondent *

*Mark only one oval.*

- 18 years
- 19-25 years
- Above 25 years

6. 3. Are you doing any part-time job? *

*Mark only one oval.*

- Yes
- No

7. 4. Are you a single parent child? *

*Mark only one oval.*

- Yes
- No

8. 5. Family Academic Qualification *

*Mark only one oval per row.*

| | Below Matric Pass | Matric Pass (10th pass) | 12th pass | Graduate | Post Graduate | Higher Qualification than Post Graduate |
|---|---|---|---|---|---|---|
| **Father/Husband Academic Qualification** | | | | | | |
| **Mother/Wife Academic Qualification** | | | | | | |

9. 6. Father/ Husband occupation *

*Mark only one oval.*

- Government employee
- Private sector employee
- Business
- Others (farmer, venders etc.)
- Retired
- None

10. 7. Total Income (Father or Husband and/or Mother or Wife) *

*Mark only one oval.*

- Below 50,000
- 50,000 to 1 lakh
- 1 lakh to 2 lakh
- 2 lakh to 5 lakh
- Above 5 lakh

11. 8. Respondent admission type *

*Mark only one oval.*

- Regular
- Private

12. 9. Are you involved in other co-curricular activities in college? *

*Mark only one oval.*

- High
- Medium
- Low

13. 10. Respondent 1st year percentage (No *round off)* *

*Mark only one oval.*

- more than 60 %
- less than 60%

14. 11. Number of children's *

*Mark only one oval.*

- none
- 1-2
- 3-5
- more than 5

15. 12. Other Information *

*Mark only one oval per row.*

| | Yes | No |
|---|---|---|
| **Are you facing any gender inequality in your surroundings (family, college, or other places)?** | | |
| **Did you face any kind of difficulty to get the transportation services from your home to college or vice versa?** | | |

16. 13. Active hour rate spending in *

*Mark only one oval per row.*

| | None | less than 1 hour | 1 to 2 hours | 2 to 4 hours | more than 4 hours |
|---|---|---|---|---|---|
| **social sites platforms ( instagram, facebook, web series, whatsapp etc.) for fun** | | | | | |
| **online video games (freefire, pubG etc.)** | | | | | |
| **other places such as canteen, café, campus etc.** | | | | | |
| **coaching classes** | | | | | |
| **self study** | | | | | |

17. 14. How much number do you give yourself out of 5 (highest 5, lowest 1) in: *

*Mark only one oval per row.*

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Confident towards pupils** | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ |
| **Emotional stability** | ⬭ | ⬭ | ⬭ | ⬭ | ⬭ |

18. 15. Attendance *

*Mark only one oval.*

- ⬭ High (more than 75 % )
- ⬭ Medium (55 % to 74 %)
- ⬭ Low ( less than 55 %)

This content is neither created nor endorsed by Google.

Google Forms